University of Zurich[UZH]

Department of Informatics

Crowdsourcing Data Analysis

Empowering non-experts to conduct data analysis

Dissertation submitted to the Faculty of Business,
Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor / Doktorin der Wissenschaften, Dr. sc.
(corresponds to Doctor of Science, PhD)

presented by
Michael Feldman
from Zurich, ZH, Switzerland

approved in September 2018

at the request of
Prof. Dr. Abraham Bernstein
Prof. Dr. Kevin Crowston

University of Zurich[UZH]

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, September 19, 2018

Chairman of the Doctoral Board: Prof. Dr. Thomas Fritz

Abstract

The development of Internet-based ecosystem has led to the emergence of alternative recruitment models which are exclusively facilitated through the internet. With Online Labor Markets (OLMs) and Crowdsourcing platforms it is possible to hire individuals online to conduct tasks and projects of different size and complexity. Crowdsourcing platforms are well-suited for simple micro-tasks which could take seconds or minutes and be completed with big number of participants working in parallel. On the other hand, OLMs are usually allowing to hire experts in flexible manner for more advanced projects that could take days, weeks or even months. Due to the flexibility of such employment models it is possible to find various experts on OLMs such as designers, lawyers, developers or engineers. However, it is relatively rare to find data scientists – experts able to preprocess analyze and make sense of data. This shortage is not surprising giving the general shortage of data science experts. Moreover, due to various reasons such as extensive education and training requirements as well as soaring demand, the projected shortage in such experts is expected to grow during the next years.

In this dissertation we explored how the crowdsourcing approach could be leveraged to support data science projects. In particular, we presented three use cases where crowds and freelancers with different expertise levels could be involved to support data science projects. We conventionally classified crowds into low, intermediate, and high levels of expertise in data analysis and proposed use cases where every group might contribute through crowdsourcing setting.

In the first case study we presented an approach of how crowds could be engaged in the review process of the statistical assumptions in scientific publications. When researchers use statistical methods in scientific manuscripts these methods are often valid only if their underlying assumptions are met. If these assumptions are compromised, then the validity of the results is questionable. We presented an approach based on micro-tasking with laymen crowds that reach quality similar to expert-based review. We then conducted longitudinal analysis of CHI conference proceedings to evaluate the dynamics of standards on statistical reporting throughout the years. Finally, we compared CHI proceedings with 5 top journals in the field of medicine, management, and psychology to compare the reporting of statistical assumptions across disciplines.

Our second case study addressed the freelancers with intermediate expertise in data analysis. To better understand what the skills that intermediate experts possess are, we conducted an interview with data scientist experts whom we asked what kind of tasks could be outsourced to non-experts. Additionally, we conducted a survey in most prominent OLMs to better understand the skills of freelancers active in data analysis. The conclusions of this study were twofold: 1) conservatively individuals with certain coding skills could be helpful in data science projects if integrated properly and 2) data preprocessing tasks are by far the biggest bottle neck activity that could be outsourced, if the coordination between involved parties is managed properly. Departing from these results, we conducted a study, where we designed a proof-of-concept for a platform that facilitated a number of experiments where non-experts were collaborating with experts through offloading data preprocessing

activities. Our results suggest that the outcome achieved with mixed expertise teams are similar in quality and cheaper than the work of experts.

Our last use case was not as much directed to alleviate the shortage in data scientists as to take advantage of the crowdsourcing setting to address inherent vulnerability of data-driven analysis. Recently, there has been a discussion among data analysis experts and researchers regarding the subjectivity of data driven analysis outputs. Namely, it has been shown that when data analysts perform data analysis where they are provided with the same data and the same hypothesis, within an NHST (Null Hypothesis Significance Testing) approach, they often reach cardinally different results. Therefore, we conducted a study where we provided 47 experts with the same data and hypotheses to answer. Through especially designed platform we were able to elicit the rational for every decision made throughout data analysis. This fine-grained data allowed us to conduct a qualitative analysis where we explored the underlying factors leading to the variability of data analysis results.

The case studies combined together provide an overview of how the discipline of data science could benefit from the crowdsourcing approach. We hope that the solutions proposed in this dissertation will contribute to the discussion on how to reduce the entry barrier for laymen to participate in data driven research as well as how to improve the transparency of how the results were reached.

Zusammenfassung

Die Entwicklung eines internetbasierten Ökosystems hat zur Entstehung alternativer Rekrutierungsmodelle geführt, die ausschließlich über das Internet ermöglicht werden. Mit Online-Arbeitsmärkten (OLMs) und Crowdsourcing-Plattformen ist es möglich, Einzelpersonen online einzustellen, um Aufgaben und Projekte unterschiedlicher Größe und Komplexität durchzuführen. Crowdsourcing-Plattformen eignen sich gut für einfache Mikro-Aufgaben, die Sekunden oder Minuten in Anspruch nehmen und mit einer großen Anzahl von parallel arbeitenden Teilnehmern erledigt werden können. Auf der anderen Seite erlauben OLMs in der Regel die flexible Einstellung von Experten für fortgeschrittenere Projekte, die Tage, Wochen oder sogar Monate dauern können. Aufgrund der Flexibilität solcher Beschäftigungsmodelle ist es möglich, verschiedene Experten für OLMs zu finden, wie z.B. Designer, Juristen, Entwickler oder Ingenieure. Allerdings ist es relativ selten, dass Datenwissenschaftler - Experten, die in der Lage sind, Analysen vorzuverarbeiten und Daten sinnvoll aufzubereiten - gefunden werden. Dieser Mangel ist nicht verwunderlich, wenn man den allgemeinen Mangel an Datenwissenschaftlern betrachtet. Darüber hinaus wird der prognostizierte Fachkräftemangel in den nächsten Jahren aus verschiedenen Gründen, wie z.B. einem hohen Aus- und Weiterbildungsbedarf sowie einer stark steigenden Nachfrage, weiter zunehmen.

In dieser Dissertation haben wir untersucht, wie der Crowdsourcing-Ansatz zur Unterstützung von Data-Science-Projekten genutzt werden kann. Insbesondere stellten wir drei Anwendungsfälle vor, in denen Crowds und Freiberufler mit unterschiedlichem Know-how involviert werden konnten, um Data-Science-Projekte zu unterstützen. Wir klassifizierten die Crowds konventionell in niedrige, mittlere und hohe Niveaus der Expertise in der Datenanalyse und in den vorgeschlagenen Anwendungsfällen, in denen jede Gruppe durch Crowdsourcing einen Beitrag leisten könnte.

In der ersten Fallstudie haben wir einen Ansatz vorgestellt, wie Menschenmengen in den Überprüfungsprozess von statistischen Annahmen in wissenschaftlichen Publikationen einbezogen werden können. Wenn Forscher statistische Methoden in wissenschaftlichen Manuskripten verwenden, sind diese Methoden oft nur dann anwendbar, wenn die zugrundeliegenden Annahmen erfüllt sind. Wenn diese Annahmen in Frage gestellt werden, ist die Validität der Ergebnisse fraglich. Wir stellten einen Ansatz vor, der auf Mikro-Tasking mit Laienmassen basiert, die eine Qualität erreichen, die derjenigen von Experten ähnelt. Anschließend führten wir eine Längsschnittanalyse der CHI-Konferenzberichte durch, um die Dynamik der Standards für die statistische Berichterstattung über die Jahre hinweg zu evaluieren. Schließlich haben wir CHI-Verfahren mit 5 Top-Journalen aus den Bereichen Medizin, Management und Psychologie verglichen, um die Berichterstattung über statistische Annahmen disziplinübergreifend zu vergleichen.

Unsere zweite Fallstudie richtete sich an Freiberufler mit mittlerer Expertise in der Datenanalyse. Um besser zu verstehen, welche Fähigkeiten die Intermediate-Experten besitzen, führten wir ein Interview mit Datenwissenschaftlern, die wir fragten, welche Art von Aufgaben an Nicht-Experten ausgelagert werden könnten. Zusätzlich haben wir eine Umfrage in den bekanntesten OLMs durchgeführt, um die

Fähigkeiten von Freiberuflern, die in der Datenanalyse tätig sind, besser zu verstehen. Die Schlussfolgerungen dieser Studie waren zweifach: 1) konservative Individuen mit bestimmten Codierfähigkeiten könnten in datenwissenschaftlichen Projekten hilfreich sein, wenn sie richtig integriert werden, und 2) Datenvorverarbeitungsaufgaben sind bei weitem die größte Engpassaktivität, die ausgelagert werden könnte, wenn die Koordination zwischen den beteiligten Parteien richtig gehandhabt wird. Ausgehend von diesen Ergebnissen führten wir eine Studie durch, in der wir einen Proof-of-Concept für eine Plattform entwarfen, die eine Reihe von Experimenten ermöglichte, bei denen Nicht-Experten mit Experten zusammenarbeiteten, indem sie Datenvorverarbeitungsaktivitäten abluden. Unsere Ergebnisse deuten darauf hin, dass die Ergebnisse, die mit gemischten Expertenteams erzielt werden, qualitativ vergleichbar und kostengünstiger sind als die Arbeit von Experten.

Unser letzter Anwendungsfall war nicht so sehr darauf ausgerichtet, den Mangel an Datenwissenschaftlern zu lindern, als vielmehr die Vorteile der Crowdsourcing-Einstellung zu nutzen, um die inhärente Verwundbarkeit datengetriebener Analysen zu beheben. In jüngster Zeit gab es eine Diskussion unter Datenanalyseexperten und Forschern über die Subjektivität datengetriebener Analyseausgaben. Es hat sich nämlich gezeigt, dass Datenanalytiker bei der Durchführung von Datenanalysen, bei denen sie mit denselben Daten und derselben Hypothese versorgt werden, im Rahmen eines NHST-Ansatzes ((Null Hypothesis Significance Testing)) oft kardinal unterschiedliche Ergebnisse erzielen. Deshalb haben wir eine Studie durchgeführt, in der wir 47 Experten mit den gleichen Daten und Hypothesen versorgt haben. Durch eine speziell entwickelte Plattform konnten wir bei jeder Entscheidung, die während der Datenanalyse getroffen wurde, das Rationale herausfinden. Diese feinkörnigen Daten ermöglichten es uns, eine qualitative Analyse durchzuführen, bei der wir die zugrundeliegenden Faktoren, die zur Variabilität der Ergebnisse der Datenanalyse führen, untersuchten.

Die kombinierten Fallstudien geben einen Überblick darüber, wie die Disziplin der Datenwissenschaft vom Crowdsourcing-Ansatz profitieren könnte. Wir hoffen, dass die in dieser Dissertation vorgeschlagenen Lösungen einen Beitrag zur Diskussion darüber leisten, wie die Eintrittsbarriere für Laien zur Teilnahme an datengetriebener Forschung verringert werden kann und wie die Transparenz der Ergebnisse verbessert werden kann.

Acknowledgements

"The capacity to learn is a gift; the ability to learn is a skill; the willingness to learn is a choice" – Brian Herbert

In the illustrated guides for fresh Ph.D. students the knowledge of a fresh doctorate student is often showed with circle whereas a tiny area bounded between a center of the circle and two adjacent radii represents the person's current knowledge as compared to the much bigger area of the circle. The idea behind this tiny circular sector, is to illustrate how little knowledge the person has gained so far compared to the available combined knowledge of humanity. When a graduate student successfully completes her PhD studies, as this illustration assures, this small area expands by a tiny pinch. If you have succeeded in doing so, the purpose of doctoral thesis has been achieved.

Today, in retrospect, I can say that this illustration is incomplete. If I had to imagine an illustration of my own studies, I would change the circle into a sphere. This would better illustrate not only to the amount, but also to the variety of the knowledge I gained. The ability to conduct independent research, identify promising research directions, set goals as well as the ability to outline the research plan requires skills that go far beyond traditional learning. Even more importantly, the ability to fail and improve accordingly, is probably the most important of all. Therefore, in addition to gaining knowledge, the last four years have been an opportunity for me to sharpen my personal skills and to gain experience in conducting my own independent research. Throughout this process there were many who supported me and made the process much more interesting and insightful.

My advisor Prof. Avi Bernstein who has been a partner for this journey for the past four years and always has shown openness and enthusiasm to the ideas and suggestions I had throughout my studies. His readiness to allow experimentation of new ideas and tolerance to failure, as well as the ability to quickly pick a raw idea and distill it into systematic research have helped me in an extraordinary way to improve my analytic thinking. His willingness to lend a sympathetic ear and to advise made him not only an academic advisor, but also a friend.

I want to thank my eternal office mate, Patrick, for sharing most of the Ph.D. journey with lots of discussions, conference trips and fun. What started as a collaboration on research projects became a solid friendship. I would also like to thank the members of my research group for sharing this journey with me and friendship: Bibek, Cosmin, Cristina, Daniel, Daniele, Ela, Juk, Lorenz, Mark, Matthias, Pengchen, Philipp, Shen, Suzanne, Thomas and Tobi. Thanks to my research collaborators with whom I worked over the years as well as to the freelancers and crowd workers which were at the focus of my research and without whom this research would not have been possible. I want to thank my committee members for the time and effort they invested in reading and judging this dissertation.

Last but not least, I would like to thank to my family whose unwavering support even in difficult times strengthened me greatly. My father, mother (R.I.P.) and my brother laid the foundations that enabled me to pursue to doctoral studies. If it were not for them, I would never reach this point. Most importantly, my wife and best

friend, Yasmin, whose love and support have no limits. Thank you for who you are and what you mean to me.

Table of Contents

Part I, Synopsis

Synopsis

Introduction

The last two decades of technological development have brought with them the data revolution. Due to significantly reduced costs associated with data creation and storage, unprecedented amounts of data are now available in various organizations. However, the availability of data is growing faster than the availability of experts with the relevant skillset to interpreting it. The demand for data scientists — experts able to provide comprehensive data-driven solutions (Davenport and Patil 2012) – currently vastly exceeds the supply of fitting graduates from the universities. Finding experts is especially challenging due to the growing number of skills needed (e.g. statistics, database structures, visual recognition, big data, or distributed computing). The growing expectation from data scientists to specialize on multiple fields has made their education even more challenging and further contributed to the shortage in data scientists (Ransbotham et al. 2015).

The need to specialize in many areas is reminiscent of other, more mature areas, in which the growing expectation of in-depth knowledge in various fields led researchers to specialize more and more on sub-fields. For example, on the outset physicists and chemists were often generalists who conducted research in various different areas of their disciplines. However, the need for extensive knowledge led these fields to hyper-specialization where researchers work together to complement their expertise (Seitz et al. 2000). Therefore, it would be natural to expect a similar fate for data scientists, who cannot sustain the demand for the scope of required knowledge. Moreover, in the light of scarce talent and growing demand, a reasonable solution would be to combine people with different levels of expertise to make it easier for data scientist experts to focus only the most challenging tasks in data analysis. Would such proposition be feasible, it will ease the need for skilled data scientists by offloading certain tasks to non-experts.

In my dissertation, I turn to the crowdsourcing approach as a potential auxiliary to mitigate the shortage in data scientists and as a source for non-experts who might be engaged in data analysis. Specifically, I propose different scenarios in which crowd-workers could assist data scientists or statistics experts. Crowdsourcing has raised interest in both the scientific and industrial community as an online collaboration model. It enables the general public to join forces to solve highly challenging problems by working together (e.g. Van Dijck & Nieborg 2009; Introne et al. 2013). Whereas the effectiveness of crowdsourcing in solving tedious, aggregative tasks is widely acknowledged, the understanding of how to crowdsource highly complex and ill-defined tasks is not yet fully discerned (Kittur et al. 2013). To date, the Information Systems discipline offers theoretical support for *collaborative work (i.e.*

work accomplished together by multiple people) and addresses the phenomenon of *crowdsourcing*. However, there are very few examples that practically couple these for solving complex problems such as data analysis.

Different online labor markets (OLMs) emerged during the last years to supply the demand for crowd-workers. For example, Amazon MTurk[1] is a platform for individuals to perform tasks that are challenging for computers but can be collectively accomplished in short time by many people. The typical task anticipated by MTurk follows a Taylorist approach where the task can be easily decomposed and the results of every crowd-worker accumulated in a straight forward manner with clear requirements, constraints, and timeframe. For medium-term tasks, which are more cognitively evolving and require higher degree of creativity, online freelance markets such as Upwork or Fiverr are more suitable. These OLMs promote projects that require high level of expertise and diverse talents. As opposed to short-term (i.e. duration of several minutes) markets like MTurk, these platforms appeal to the mid-term (i.e. days or weeks) dynamic employment structure, where experts can be recruited and released in an agile manner based on the ad-hoc needs. Differently to the micro-tasking OLMs usually support more individualistic type of work, where one freelancer is responsible for the whole project. Due to the market need to hire ad-hoc talent, online freelance markets have significantly evolved during the last years and now bring together millions of freelancers and employers (Feldman et al. 2017).

In line with the general shortage of data scientists, it is fairly uncommon to find data analysis *experts* on crowdsourcing or OLM platforms. The demand for such experts in western world is high enough to quickly absorb them into traditional, permanent, jobs. On the other hand, in developing countries freelancing is much more attractive due to the opportunity to earn global salary without geographically relocating. However, in order to earn an expertise in data science one has either to go through a long track of self-education or to study at one of the leading universities of the region. As a result, the presence of data science experts on the online labor markets is insufficient.

It is, however, not unusual to find workers possessing some partial knowledge and/or willing to learn a new topic. These workers might be not able to carry out a full, end-to-end data science project, but would be instrumental in supporting data scientists in their endeavor. For example, data science project might start with identifying the relevant data sources for the analysis. This task involves expert judgment since irrelevant or noisy data will undermine the results. On the other hand, activities associated with data integration and cleaning could be done by not very skillful freelancer if the right safeguards are applied. Moreover, some tasks are simple enough to be executed by laymen crowds – individuals without any

[1] www.mturk.com

expertise relevant to data science. For instance, OLMs like CrowdFlower[2] are engaging crowds in manual data cleaning and preprocessing resulting in data which is used for further statistical modeling.

Lastly, while expert data scientists usually are not present on OLMs, they still use crowdsource-based platforms to communicate with each other. Specifically, they often participate in specialized groups where they can discuss the latest advancements in data science and post an open question regarding the challenges they face during their work. For example, websites such as StackOverflow or specially dedicated Linkedin and Facebook groups host such discussions where data scientists exchange their opinions and help each other with advanced topics of data analysis. Therefore, employers often use these platforms as a recruiting channel to get in touch with experienced data scientists.

To summarize, crowds with different level of expertise present on different crowd-based platforms: laymen on the micro-tasking crowdsourcing platforms, freelancers on OLMs and experts on the especially dedicated groups and forums. We therefore ask a central question about the contribution of crowdsourcing to data science:

How can crowdsourcing help mitigate the increasing shortage of data scientists?

We answer this question by examining crowds with different levels of expertise. Laymen crowds – with no expertise, who are mostly active on micro-task labor markets. So-called non-experts – freelancers with some Software Engineering experience but no extensive training in data science, who are mostly present on OLMs, and experts – experienced data scientists and statisticians with advanced relevant education and rich experience, who are mostly present on expert forums and dedicated mailing lists.

[2] https://www.crowdflower.com/

Problem statement

Data science projects are highly iterative and unstructured and usually require from data scientist cognitive effort and broad set of skills (Bernstein 2000). Throughout data analysis, data scientist ought to make decisions that are highly subjective and interpretative such as what data to include, to sample, to pre-process, and to analyse. To aggravate further, differently from other disciplines, it is extremely challenging to evaluate the quality of data analysis post-fact. The output of flawed or biased data analysis might easily look very similar to a well-performed analysis. This means that with so many degrees of freedom in decision making and challenge to evaluate the quality of the results, the experience of data scientist plays a key role.

Following this understanding, during my PhD studies, my collaborators and I explored how crowds with different level of relevant expertise can contribute to data analysis. Specifically, in three different studies we proposed scenarios of how crowds with basic, intermediate and high expertise can be instrumental in data analysis. However, by conventionally categorizing the data science expertise of crowds into low, medium and high, we faced an uncertainty regarding crowd workers with medium level of expertise. Based on preliminary study and literature review, we reached the conclusion that crowds with medium level of expertise are those who possess some coding skills but no significant data analysis experience. We call these freelancers "non-experts". Yet, it is unclear to what extent they are capable of carrying out data science projects. The existing literature does not provide us with conclusive answer as to whether there is sufficient number of such freelancers and what concrete skills do they possess. We, therefore, started our research with exploratory study which was aiming to better understand the talent available on freelance OLMs.

Depending on the level of expertise and type of market place, we propose different platforms that support task execution. For the crowds who are likely to not have any expertise on MTurk, the task has to be simple, well-specified and with easy to quantify quality assurance mechanism in place. Differently, non-experts might perform certain tasks but have to be closely guided by the data scientists. The communication with data scientist in charge should be on the one hand flexible, to allow for a rapid feedback loop, but on the other hand, not too intense. A very demanding communication burden would outweigh the benefits resulting from outsourcing some parts of data analysis to non-experts. Yet, for experts, the communication system is less crucial as they usually perform analysis independently. Still, data science experts face another type of challenges. One of such challenges that gained attention during the last years is the ability to reproduce data science projects and thus assure the reliability of data analysis outputs. Among the underlying reasons for difficulty to reproduce data analysis, is a sequence of implicit decisions that data scientists make as they perform data analysis. Therefore, it is important to elicit such decisions and provide a means to communicate them in a transparent way.

This manuscript describes three studies proposing how crowd workers with different levels expertise might be involved in Data Science projects. In the first study, we developed a setting, where laymen crowd workers reliably evaluate the quality of statistical assumptions reporting in research papers. Such assumptions are critical to assure the validity of statistical method and, if violated damage the credibility of the results.

Our second study revolved along non-experts available on freelance online labour markets. Specifically, we realised the opportunity in outsourcing data preprocessing tasks – activities associated with data cleaning and transformation such that the subsequent analysis could be executed. These tasks considered as bottle neck activities of every data analysis project and consume up to 80% of data scientists' time (Gutierrez 2015). Provided sufficient oversight by a data science expert, non-experts should be enabled to efficiently preproccess the data and thus offload a significant burden from expert data scientists. However, to be able to do so, data scientists and non-experts should be able to collaborate efficiently. We therefore followed a design science approach to explore the necessary features required for efficient data analysis execution with mixed expertise crowds.

Lastly, we leveraged the crowdsourcing approach to research the variability in data analysis among experts. Specifically, we explored the underlying factors which are causing different experts to reach varying results when they analyze the same data (Silberzahn & Uhlmann 2015). This problem drew attention in the recent years in scientific community and is central to the reliability of the scientific results reached through data analysis as such. Usually data-driven research is conducted by one research group. We crowdsource data analysis project where a large number of researchers perform data analysis in parallel. We then analyze the results to identify the reasons leading to different results. To do so, we designed an experiment where many experts were presented with the same set of hypotheses and asked to conduct data analysis using identical data. By using especially designated web-platform, we recorded all implicit decisions made by analysts during analysis and identified potential factors leading to different results.

Terminology

Throughout this thesis we use the following terminology:

Cognitive data analysis:	(Grolemund & Wickham 2014) define it as a scientific model to explain the data analysis process, which attempts to create understanding from data.
Non-experts/Lay statistician:	in our context are individuals that possess basic or intermediate coding skills and no or very basic data analysis skills.
Preprocessing:	Broadly, refers to all activities applied to data before the data can be used for statistical modeling
Online Labor Markets	When we refer to Online Labor Markets we mean freelancing web platforms where freelancers might be hired all around the globe to accomplish projects in variousdomains such as software engineering, design, prrofreading and editing.
NHST:	Null Hypothesis Significance Testing is the common statistical inference framework that includes parametric methods. These methods are fully valid only when the underlying statistical assumptions are met.[3]

Research Questions and Hypotheses

We explored the topic of crowdsourced data analysis in four steps translating into their respective research questions:

1. First, we report on a case study, where laymen crowd workers review the quality of statistical assumptions reported in research papers.
2. Second, lacking the sufficient literature on the available data science talent on OLMs, we conducted an exploratory study to better understand the qualifications of available data science freelancers.
3. Third, we report on a case study, where we develop a proof of concept to show the feasibility of crowdsourcing data preprocessing to non-experts.
4. Fourth, we report on a study, where we crowdsource data analysis among experts in order to explore factors driving variability in data analysis.

[3] Note that some statistical methods are robust to assumption violations, however referring to this alleviation has to be rooted in relevant literature.

Our first research question seeks to understand how crowd workers without any expertise in the field of statistics can contribute to statistical validity of scientific publications via short term micro-tasks.

Various scientific disciplines often rely on data analysis to produce knowledge. This assumes that the underlying data analysis is following a methodology that allows to review the results as statistically valid and reliable. To assure statistical validity, as well as other aspects if research quality, research manuscripts are reviewed by scientific peer reviewers prior to publication. This process is the major safeguard applied by the scientific community to assure that the published papers adhere to high standards and result from rigorous scientific process. However, the peer-review system applied by most of the scientific outlets imposes a substantial burden on the reviewing researchers. In addition to conducting their own research, teaching as well as administrative responsibilities, scientists are also tasked with reviewing papers of their colleagues. To aggravate further, evaluating the quality of statistical inquiry is very challenging due to limited information provided in papers. Authors often partially omit the information describing statistical methodology they follow, due to the rigorous space limitations imposed by most of the outlets. Moreover, supplementary materials to the papers are either not provided or lack sufficient information. Lastly, sometimes researchers are experts in their field but not trained enough in statistics to judge about the nuances of statistical methodology described in the manuscript.

It is not surprising, therefore, that the scientific community has expressed a great concern regarding the quality of published research (e.g. Klein et al. 2014). The concerns are raised across different disciplines such as medicine, biology or psychology and address various aspect of research validity and reliability such as significance testing (e.g. a discussion around p-value), data dredging, biases, confounding, method selection and testing for assumptions (Davey Smith & Ebrahim 2002). Most of these problems would have been solved had it been possible to allocate sufficient resources for a rigorous examination of each manuscript.

Thus, our first research question is revolving around how to support a review of statistical validity in research papers. We limit our scope to one aspect of statistical validity. Specifically, we are exploring *how to engage laymen crowd workers in reviewing the statistical assumption reporting in research papers*. Statistical assumptions reflect the characteristics of the data. If these assumptions are violated, the conclusion of the research and its interpretation is called into question. We, therefore, designed a web-based tool where crowds are leveraged to evaluate the quality of statistical assumptions reporting in research papers. Our goal is to find a reliable way to outsource part of the review process to crowd workers without compromising on the quality of the review as it would be reached by expert peer reviewers.

RQ1: How can crowd workers evaluate the quality of statistical assumptions reporting in research papers?

Statistical validity plays a key role in how well study results are accepted. The violations of statistical validity often invalidate the paper and, if published, sometimes even causes retraction (for few such examples see RetractionWatch[4]). However, lack of reporting of statistical assumptions is not necessarily due to lack of testing, but rather might be a result of authors' assumption that it is too obvious to be mentioned that the assumptions were tested. Moreover, sometimes the assumption testing is conditioned on other parameters such as sample size or data source. It is therefore not possible to elicit all nuances neither by crowds nor by expert reviewer without a thorough investigation. Yet, transparent statistical assumption reporting is advantageous for both reviewers and readers. Involvement of crowds will ease the burden of the review and empower public participation in scientific research. This is hopefully to result in a more transparent statistical reporting in research publications.

We propose an automated way to evaluate reporting of statistical assumptions using crowd workers with no statistical knowledge. By automating statistical assumption reporting evaluation at scale and at low cost, we assist researchers to detect potential flaws related to the reporting. Our approach might be used by authors of research papers as means to identify potentially missing discussions about assumptions and to account for them in their manuscript prior to paper submission. It can also be used by editors for initial paper screening to identify missing information. Lastly, such tool can benefit the reviewers to quickly evaluate potentially missing assumption reporting.

> H1.1: Laymen crowd workers might contribute to the review process of the statistical assumption reporting of the research manuscripts

This hypothesis is explained with an experiment, where we evaluate 131 research papers from a leading computer science conference and compare the results with the ground truth data. The ground truth data was generated by three experts manually annotating 100 papers. Using pattern matching, we identify statistical methods and their corresponding assumptions. Crowd workers presented with text snippets that contain highlighted method and assumptions and tasked, through number of questions, to evaluate whether these two terms are semantically related. In other words, they are asked if the highlighted assumption indeed relates to the method. Crowds answers are aggregated using quality assurance metrics common in crowdsourcing (e.g. Beat-by-K). We then evaluate the results by comparing them with the ground truth data manually annotated by experts.

The hypothesis is translated into a binary classification problem where the aggregated output of crowd workers for every method – assumption occurrence either match or not to the ground truth. The validity is assessed through common metrics of binary classifier such as accuracy, precision, recall and F1 measures.

Having such tool enables unique opportunity to analyze the quality of statistical assumption reporting in big corpus of scientific papers. We therefore draw on this to

[4] http://retractionwatch.com/

run a cross-disciplinary study to compare the quality of reporting on statistical assumptions in leading journals (measured by impact factor) in medicine, management, psychology and computer science. This provides us a unique opportunity for big scale comparison of hundreds of papers from six leading venues across different fields. Therefore, the second hypothesis we investigate is related to the quality of the statistical reporting across journals

> H1.2: The quality of the reporting on the statistical assumptions is similar across different fields

In total we conducted a large-scale analysis of 442 papers where we analysed the assumptions of five common statistical methods: ANOVA, t-test, Linear regression, Chi-Square, and Logistic regression. The list of statistical assumptions was finalized based on the literature (Field 2013) and in collaboration with statisticians. Our major goal was to answer three questions: (1) do fields differ in which methods they use, (2) do fields differ in which assumptions they report, and (3) do fields differ in the extent to which assumptions are reported. In this study we did not aim to prove a statistically significant differences between the research fields as it would require a much bigger sample size with papers from many more journals. Our goal was to provide initial evidence of whether there are differences in statistical assumptions reporting between journals coming from different disciplines and to point to potential future research in this direction.

We so far presented a case study demonstrating how laymen crowd-workers can be instrumental for the scientific data-driven analysis. Our next research question is dealing with freelancers who have some coding skills but no advanced data science training and experience. While we envision the non-experts to possess certain coding skills, it is actually unclear whether this type of workers is available on OLMs. There are some studies describing the talent available on OLMs (e.g. Kalleberg & Dunn 2016), however there are no studies describing the available talent in software and data science in detail. Moreover, it is unclear what data analysis tasks could be potentially crowdsourced to freelancers. Specifically, we are interested not in the full end-to-end project outsourcing, but in the tasks that data scientist could outsource to freelancers to speed up internally executed projects. Lastly, to complete an overview on this topic, we explore what are the obstacles in crowdsourcing data science (sub-) tasks to freelancers.

RQ2: How can we benefit from the existing talent available on OLMs to alleviate the need for data scientists?

To understand what are the tasks that expert data scientists would consider outsourcing, we conducted interviews with data scientists in Germany and Switzerland. Our interviewees either hold positions of data scientists/data analysts or are primarily occupied with data analysis in their daily work. These interviews helped us to better understand what the tasks are that data scientist would be willing to outsource. We further asked data scientists to describe the major obstacles for outsourcing. After analysing all interview transcripts, we developed an online questionnaire based on the results of the conducted interviews. The purpose of the questionnaire was to study the distribution of skills, expertise, and knowledge in the

population of the most prominent freelance platforms. The answers from 80 respondents were then collected and qualitatively analyzed to identify the most in-demand skills. This research therefore tests the following hypothesis:

H2.1: There exists a talent on OLM that can contribute to data science projects

We explore this hypothesis by statistically evaluating the skills found on OLMs. Specifically, the final questionnaire primarily consisted of 5-point Likert scale questions, which were designed to capture freelancers' skills expertise and knowledge of major statistical tools (for more detailed explanation of the methodology please see page 73). Similar to other papers following mixed-method approach (e.g. Schlauderer & Overhage 2013), we analyzed whether the self-reported skills significantly different from the average skills to be expected. Following literature guidelines (de Winter & Dodou 2010), we performed one-sample two-tailed t-tests. This allowed us to identify data analysis skills prevailing among freelancers available on OLMs. Note, that the skills in our study are self-reported and prone to be positively biased. We therefore took a conservative stance while reviewing self-reported skills. Despite this limitation, the results still present strong evidence that the talent available online is sufficient. It is not fully clear to what extent the statistical and machine learning knowledge of freelancer is advanced. *Yet, based on their education and past projects, they clearly possess at least certain coding skills*.

Moreover, most of the data scientist mentioned *data preprocessing as a task that they would be happy to outsource*. This is not surprising as data preprocessing is by far the most time-consuming part of data analysis and it is known to be a bottle-neck activity of data analysis projects. On the other hand, as reported by interviews, once it is well understood how to preprocess the data, these activities require no advanced data science knowledge. The major reason for avoiding outsourcing these tasks however was the concern about the privacy and trust in freelancers. The communication overhead was another major obstacle hampering crowdsourcing data preprocessing. Would these obstacles alleviated, the shortage in data scientists could be eased by an efficient work environment where experts and non-experts coordinate and distribute the task according to their complexity and requirements.

The realization that non-experts might be useful in data preprocessing, led us to our third research question. Having a heterogeneous team of data scientists, data engineers, senior and junior software developers within a company's premises is a common arrangement. It is also very common to have geographically distributed teams of software developers working together. However, would it be possible to support teams of data scientists and non-experts to work together outside any organizational structure tied only by a short-term contract on OLM? More specifically, *would it be possible to organize these different crowds to collaborate together, such that non-experts could offload the data preprocessing from experts?* Apparently, such collaboration is conditioned on suitable environment that will enable efficient coordination such that the burden of (micro-) managing non-experts will not

outweigh the benefits of outsourcing this task.

We therefore explore how non-experts can be involved in data analysis, and especially in data preprocessing. Following the design science approach (Peffers et al. 2008) we iteratively develop a web-prototype that will facilitate the data analysis through iterative refinement of results up to its successful completion. Our major goal is to build a proof-of-concept prototype to showcase how non-experts can effectively collaborate with data scientist in crowdsourcing setting.

RQ3: How can we outsource some parts of data analysis to freelancers with some coding skills but no advanced data science experience?

The goal of the prototype was to enable the coordination of diverse crowds and facilitate task execution upon its successful completion. Specifically, experts should be able to split the task in a straightforward way, to delegate the task to crowd workers and oversee the execution of tasks (Langlois 2002). Since data analysis is an iterative process, which is in this case executed with non-permanently hired workers, the prototype was designed to keep track of the iterations and manage the project even when the crowds or the manager are changing. Note that there are other platforms like GitHub or Trello that can fulfil similar requirements. However, since we assume non-experts to take part in data analysis we aimed at straightforward environment which can be easily adapted by novices. Moreover, we wanted to create a platform where the task is not only decomposed and managed, but also online executed and can be easily accessed by the person holding the manager-role.

The decomposition was relatively simple and guided by data scientists themselves. We evaluated whether the decomposition lead to reduced complexity through a user study by answering the following hypothesis:

H3.1: It is possible to decompose typical data analysis projects into small enough tasks, reducing their complexity to a level where non-experts can accomplish them

To test this hypothesis, we conducted four data analysis projects where we ascertained whether it is possible to decompose the selected data analysis projects into sub-tasks such that the complexity of the sub-tasks is reduced compared to the overall complexity of the project. We evaluated this qualitatively with a survey, where we asked crowd workers to report on the perceived complexity of the project and the corresponding sub-task (more on this can be found on page 101). Hence, we were able to confirm H3.1 through the respondents' ratings of the corresponding survey questions.

Second, to be useful, the proposed solution has to be comparable in quality to traditional expert-based data science projects. Hence, to answer whether the proposed method is feasible and can reach the desired output of collaborative data analysis with mixed-level expertise teams, we proposed the following hypothesis:

H3.2: The quality of the results produced by a team of non-experts is comparable to the one achieved by experts

We tested this hypothesis by statistically comparing the results of the projects performed by experts with the results of non-experts who used our platform. As the data analysis projects, we used for evaluation are publicly available on Kaggle,[5] we explicitly asked the participants not to search and browse for the solutions. We also compared the code and the solutions' logic to assure that the code has not been inspired by the original solution. We attempted to cover a range of typical data analysis projects with complexity that meets real-world scenarios.

We selected the projects based on the following criteria: a) the projects should be implemented in either R or Python, since these are the most popular languages in data analysis, b) the projects should contain a relatively large preprocessing part, c) the projects should encompass various types of data analysis such as descriptive statistics, visualization and prediction, d) the projects should be conducted by individuals that can be considered as experts either based on their verified biography or because of their high ranking on Kaggle, and e) as the projects have to non-trivial, we limited the minimal size of the project to be about 150 lines of code (implemented by experts) as well as chose projects with significant number of up-votes and history of comments such that it can be assumed that the code went through a substantial public review.[6]

So far, we discussed how laymen crowds and non-experts can contribute to data science in crowdsourcing setting. The last research question revolves around how experts, when organized in crowdsourcing setting, can contribute to (the science of studying) data science. We found an answer to this question when we looked on a matter which in recent years draws a growing attention within the scientific community: the variability in data analysis results. In a typical data-driven research project, authors draw conclusions based on the analysis of the data. If published, the results add to the body of knowledge and other researchers then draw from these conclusions in their research. Interestingly, the results are not as objective as they are usually perceived. In fact, there are often numerous analytic strategies that could be used on the same data. Variations in such strategies could produce very different results - the so-called "garden of the forking paths" (Gelman & Loken 2014a). This means that even when analysis is performed by many experts using the same data and pre-defined hypotheses, they often reach fundamentally different results (e.g. see study of Silberzahn & Uhlmann 2015). Crowdsourcing offers just the right toolset to inspect this question at scale. We therefore leverage the crowdsourcing paradigm to explore what are the behavioural factors underlying the variability in data analysis results.

RQ4: What are the factors potentially causing variability in data analysis results?

[5] www.kaggle.com

[6] This paragraph is cited from the third chapter in this thesis, *"Towards Collaborative Data Analysis with Diverse Crowds – a design science approach"*, page 86

To explore possible forking paths in data-driven research, we crowdsourced a data analysis project to many expert analysts. By doing so, we were able to observe the roadmap of different analytical alternatives and decisions. We therefore analysed the steps undertaken by data analysts and explored factors underlying the implicit decisions made throughout a data analysis. The crowdsourcing approach we adapted in this study is very instrumental, as it allowed many analysts to test the same hypothesis on the same complex dataset, while being blind to each others' analytic approaches. When similar results are obtained by many analysts, scientists can speak with one voice on an issue. Contrarily, if the observed effects are highly contingent on subjective analytic decisions, the results are called into question. Different analysis strategies may converge to very similar estimated effects, indicating robustness in results despite variation in analysis strategies. Alternatively, the estimated effect may be highly contingent on analysis strategies. If so, then subjectivity in applying scientific methods and ambiguity of scientific results is made transparent. In this latter case, a crowdsourcing approach offers a unique opportunity to identify sources of variation in data analysis.

H4: There exist cognitive and behavioral factors driving variation in data analysis

To address this hypothesis, we designed an online platform to trace the decisions experts do during data analysis. The platform is based on Rstudio server[7], a popular data analysis platform that allows users to conduct analysis remotely based on the familiar Rstudio interface. Note, we could not reuse the platform developed for H3, because we are now targeting users of very high expertise level which have different development environment.

We then added a number of features to better follow the workflow of analysts. For instance, once the analysts reached certain number of *executed* commands, we prompted them to explain the code. Drawing from rational design research, analysts were asked to explain the goal, reason and the alternatives of the relevant code. This way we were able to abstract executed code into semantic blocks explaining the rational and the considered alternatives of the code. When the analysis was over, analysts were presented with the semantic blocks they created during analysis and could graphically restructure them to best reflect the workflow they followed during data analysis[8]. Once the experiment was over, we qualitatively analysed the annotated code. Using inductive coding, we identified patterns that are varying among researchers and might lead to the variability in results. Note, that we neither could unequivocally claim that these factors are responsible for the results variability, nor could we quantify how each factor adds up to the results variance. However, we identified the behavioural patterns that varied across analysts and proposed further examination of these factors in the strict controlled experiments.

The variability in data analysis can be seen on the grander scheme as a constituent of ongoing reproducibility crisis in research. Recent attempts to reproduce research

[7] https://www.rstudio.com/products/rstudio-server/

[8] see platform demonstration here - https://goo.gl/rnpgae

across different disciplines resulted in a disturbing observation where only small fraction could be reproduced (e.g. Open Science Collaboration 2015). While there are multiple factors which are proposed as responsible for this, the way researchers analyse the data seems to be one of most central. If even with fixed hypothesis and identical data researchers reach different results, it is only logical that when data is collected separately and according to subjective judgment, the results would probably greatly vary.

Currently in research papers it is uncommon to provide a detailed reporting on the way analysis is conducted. Authors usually only describe general methodology they followed. However, the behavioural factors explored in this study under current practice remain inconspicuous. We hope that this study will raise an awareness of these problems and spark studies proposing (maybe crowd based) platforms that will allow to collectively decide on the best or, most acceptable route in the garden of forking paths of data analysis.

Contribution Summaries

The outlined research questions and their corresponding hypotheses are investigated in four distinct research projects. In this section we summarize the findings and discuss the contributions

The following contributions correspond to the research questions

RQ1: How can crowd workers evaluate the quality of statistical assumptions reporting in research papers?

RQ2: How can we benefit from the existing talent available on OLMs to alleviate the need for data scientists

RQ3: How can we outsource some parts of data analysis to freelancers with some coding skills but no advanced data science experience?

RQ4: What are the factors potentially causing variability in data analysis results?

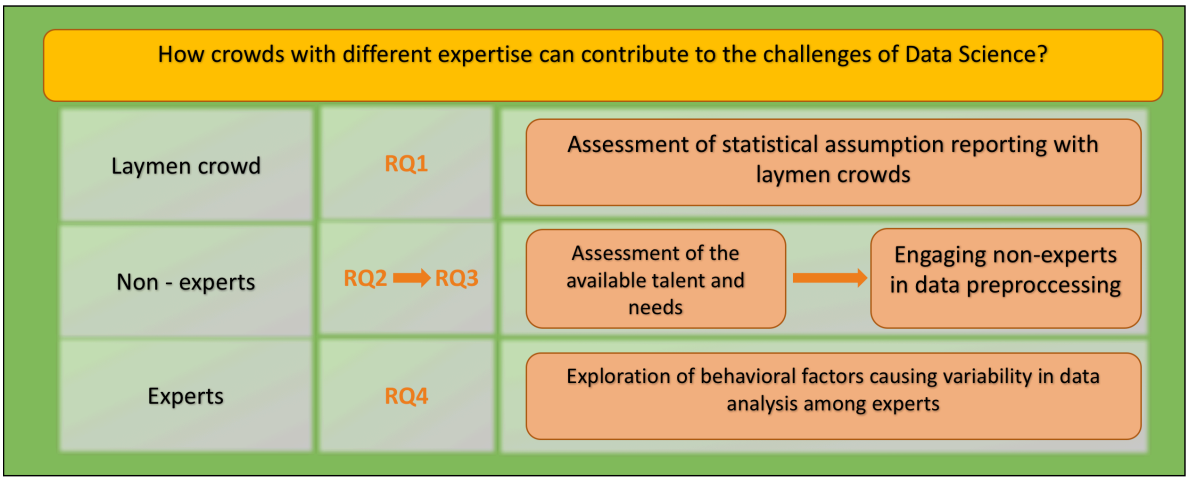

Figure 1: schematic representation of the studies constituting this dissertation

All projects address the overarching question of my thesis. They propose how crowds with different expertise can address some of the current challenges of data

science, whereas they complement each other based on the knowledge level of the people employed (see Figure 1). The first study is a use case presenting how laymen crowds could review research papers to assess the quality of statistical assumption reporting. The second stage of my dissertation evolves around non-expert freelancers and consists of two consecutive studies. First, we identify what are the skills of these individuals and then propose a setting where they can be engaged in online collaboration with experts and help to prepare data for analysis. Lastly, in the third stage, we show how crowdsourcing can be useful in identifying the underlying factors responsible for variability on data analysis results.

Q1: Case Study for crowdsourcing the evaluation of statistical reporting to laymen

The first question presents a case study exemplifying how laymen crowd workers can be instrumental in evaluating the quality of statistical assumptions reporting of research papers at scale. We developed crowdsourcing method that presents crowds with excerpts from papers and through few questions seeks to evaluate whether the mentioned statistical method and assumptions are semantically relevant. The evaluation of our approach resulted in a Precision of 82%, Recall of 83%, and F1 of 83%. This means that 82% of the assumptions identified by crowds match the ground truth and vice versa, 83% of the method assumptions existing in ground truth were identified by crowd workers. Overall, the crowd-based method performed slightly worse than expert annotators who also sometimes disagreed with inter-rater agreement of 0.79.

The results of the evaluation support Hypothesis 1.1, which suggests that laymen crowds can be instrumental in the review process of statistical assumption reporting in research publications. Using our system, it would be possible to assist experts in disemboguing whether the assumptions were taken into account by the authors.

The developed method allows to analyze the quality of statistical assumptions at scale without compromising on the reliability of the results. Therefore, in addition to CHI we analyzed five top journals across different fields to evaluate the standards on reporting of statistical assumptions for five most common statistical methods. We found a very low reporting rate of assumptions across all journals without any assumption standing out. Note however, there are different expectations for different methods depending on the journal. For example, assumptions of linear regression are much more often mentioned in journals belonging to psychology and management fields as compared to medical and CHI domains. In contrast, ANOVA assumptions are much better reported in medical journals than in other fields. These could mean that different fields have established traditions regarding the perceived importance of different assumptions.

The results of the comparative study we conducted supports the Hypothesis 1.2 assuming low rate of statistical assumption reporting across different disciplines. Even though this hypothesis was not statistically evaluated due to the challenging sample size required for cross-disciplinary comparison, the results clearly indicate similarly low reporting rates across different fields.

The contributions of this study can be summarized as follows: (1) a scalable crowd-based method to review the reporting of statistical assumptions in research

publications, and (2) cross-disciplinary analysis of 442 research papers to reveal the current standards of statistical assumptions reporting.

Q2: OLMs as a talent source that can supply the need in data scientists

The second research question addresses the need to better understand how the growing shortage in data scientists can be relieved with talent available on online labor markets. To do so, we conduct interviews with data scientists in order to understand what skills are required from freelancers to take part in data science projects. Following this scenario, in-home data scientists will be responsible for the data science projects while some tasks will be outsourced to freelancers. We further explore the major challenges to crowdsourcing data science tasks to freelancers and evaluate the skills present on major OLMs. In this study we follow a sequential mixed-method approach where we first interview 20 data scientists, subsequently learn the sought-after skills of freelancers, and then survey 80 respondents from various OLM regarding their skills and experience.

Our study suggests that in addition to technical affinity and mathematical skills, the domain knowledge, eye for esthetics, and communication skills play a key role in successful outsourcing of data analysis. As it was recognized by both data scientists and freelancers, the hurdles in outsourcing data science projects were communication issues, quality assurance, knowledge gaps, as well as privacy and confidentiality of data. Lastly, we asked data scientists what tasks they would most likely outsource to freelancers. Undoubtedly, the most desired task to be outsourced are data cleaning and preprocessing. This is due to lack of high level expertise required for most of this task and time-consuming character of this task. However, data cleaning requires understanding of the domain context and has to be done iteratively. Therefore, a coordinating effort with freelancers might be substantial where at each iteration the specifications have to be clearly conveyed – something that is not always realistic.

The results of the conducted interviews together with subsequent questionnaire, support Hypothesis 2.1 which suggests that there is a talent on online labour markets that can contribute to data science projects. This could be done by assisting data scientists in the tasks that are time-intense but do not require any advanced training in data science.

The contributions of this study can be summarized as follows: (1) we explored what skills freelancer data analysts are expected to have, (2) we explored the major hurdles for data analysis crowdsourcing and discussed potential remedies for this problem and, (3) we reviewed the skills freelancers on OLMs possess.

Q3: Crowdsourcing data preprocessing to non-experts

The third research question stems from the previous research and aims at providing a proof of concept that the bottle-neck activity of data preprocessing can be reduced in a crowdsourcing setting. Specifically, we design an online prototype that facilitates a collaborative data analysis where data analysis expert orchestrates non-experts to perform data analysis online. Using design science approach, we propose a prototype based on IPython notebook (since renamed to Jupyter Notebook) and reveal some important features for collaboration such as support for dynamic

development, code deliberation, and project journal. The proposed solution differs from existing tools for distributed software development in its scope: supporting non-experts with limited coding skills in data preprocessing tasks.

We support our hypotheses through four data analysis projects that were executed in two iterations. The results demonstrate that projects can be accomplished in economically competitive cost with freelancers available on OLMs. Most of the freelancers in our experiment were bachelor or master students working part-time as freelancers. Despite being mildly proficient in coding (self-reported 3.2 out of 5 on coding skills) and no background in data analysis, they were able to produce results statistically equivalent to the results of experts.

The Hypothesis 3.1 suggests that it is possible to decompose typical data analysis projects into small enough and less complex tasks, such that they can be accomplished by non-experts. The self-reported evaluation of task complexity collected in our study provide positive evidence that the tasks decomposed by expert could be accomplished by non-experts. Moreover, the equivalency in the results of projects performed by experts and teams with mixed expertise, support Hypothesis 3.2 that the quality of the results produced by a team of non experts is comparable to the one achieved by experts.

The contributions of this study can be summarized as follows: (1) a method for collaborative data analysis dedicated to crowds with basic coding and no data analysis skills, (2) proof of concept for the proposed method that demonstrates that our approach is both cost-effective and can produce results equivalent in their quality to those of experts, and (3) a prototype of online platform that supports collaborative data analysis of freelancers with different levels of expertise.

Q4: Exploring the sources for variability in data analysis with crowdsourcing

In our last research question, we leveraged the crowdsourcing phenomenon to answer a theoretical question regarding the variability in data analysis. Data analysis experts are required to make myriad tacit and not-obvious decisions. As a result, different experts might reach different results. Unfortunately, it is very rare in the scientific world that the same data and research questions are analyzed by more than one researcher or research group. As a result, the conclusions of an analysis is to certain extent stochastic and questionable.

To answer Hypothesis 4.1, we designed a platform which allows us to trace all decisions made by experts through data analysis. Thanks to careful recording of the decisions which experts make during their analysis, we are able to create a road map of tacit factors that are likely to impact the final results of data analysis. We did so by recording all commands executed by analysts and by asking them to explain the rational of the executed commands in an on-go manner. To avoid the burden of distracting analysts from their work too much, we carefully designed the platform to minimize the interruptions. At the end of the data analysis we also asked our participants to create visual workflows that would reflect the course of their data analysis. Using this unique data from more than forty experts, we qualitatively derived the factors driving variability in data analysis following general inductive approach.

The results of this study support Hypothesis 4.1 suggesting the existence of cognitive and behavioural factors driving variation in data analysis. While this study was dedicated to exploring potential factors responsible for the variability, future research will need to examine the impact of each of these factors.

The contributions of this study can be summarized as follows: (1) a web-based platform that supports transparent data analysis, (2) an exploratory study that of the factors contributing to the variability in data analysis, and (3) a model that is systematically conceptualizes cognitive processes underlying the variability of data analysis.

Outline

In the following, Part II presents the four articles that belong to this cumulative dissertation. First, we present a case study demonstrating how laymen crowd workers can contribute to the statistical assumptions review of research papers (page 33), followed by an exploratory study about the data analysis talent available on online labor markets (page 67). We then present a study demonstrating the feasibility of integrating online available non-expert freelancers to offload the preprocessing tasks (page 88). The last article leverages the crowdsourcing to explore a pending question in data analysis of variability and subjectivity in data analysis due to cognitive diversity of analysts (page 113).

Part III outlines the limitations of the reported findings and presents directions for future work, followed by a conclusion.

Part II, Papers

Assessing Statistical Assumption Reporting in CHI and Other Fields

This chapter is based on a paper that is currently under review of ACM Transactions on Computer-Human Interaction journal. The authors contributing to this chapter are: Michael Feldman, Patrick M. de Boer, Jen Mankoff, Carolin Strobl, Abraham Bernstein.

Assessing Statistical Assumption Reporting in CHI and Other Fields

Abstract

Researchers often use NHST statistical methods, to evaluate hypotheses. Recently, the use of such methods has been called into question. While parts of the scientific community argue for the elimination of NHST, scientific research still makes extensive use of it. As such, this article's focus is on assessing one area of NHST that is still ambiguous and not always in consensus: The reporting of assumptions underlying the employed statistical methods. Although assumption violation does not always threaten statistical validity, we argue that lack of reporting of assumption checking hampers the assessment of statistical reasoning in published works and the reusability of results. In addition, we show that neither the review process nor the outcome of that process effectively ensures reporting in this domain. This paper presents (1) a scalable method for checking for reporting violations with an F-score of 83% and a Kappa of .7, (2) an analysis of 261 papers published in CHI proceedings between 1989 and 2016 which shows that only 13% of papers using the statistical methods ANOVA or t-test reported at least one of their assumptions and the improvement over time is small, as well as (3) a comparison of a single year of CHI with top-journals from other fields such as medicine, psychology, and management showing less than 6% of the expected assumptions are reported on average (across 5 common statistical methods).

Our work calls into question, whether we as a community have been meeting our own implicit standards or those of the fields that we draw upon methodologically. By using our tool the agreed standards could potentially be verified in a transparent and accessible fashion that could benefit both authors and reviewers.

1 Introduction

The reproducibility of published scientific studies is a pressing concern, especially given the influence these studies may have on policy, medical treatment, and other real-world problems. In Nature's survey of 1,576 researchers who answered a questionnaire on reproducibility in research, nearly 90% advocate for better statistics and more robust experimental design (Baker 2016a). One of the reasons for this crisis is the abuse of null-hypothesis statistical testing (NHST) methods, which can result in threats to the statistical validity (Leek & Peng 2015). This discussion surfaced in CHI, for example, in the form of ignoring effect sizes (Kaptein & Robertson 2012). However, awareness of these issues has not necessarily translated into better practice. Attempts to switch to alternative approaches (such as Bayesian methods) have been proposed. However, the vast majority of scientific research still depends on traditional null hypothesis statistic testing and related methods (which we will refer to as NHST throughout this article). In the meantime, it is important to ensure that NHST, when used, is both valid (*applied correctly*) and validatable (*reported correctly*) in published articles. In addition, reporting of assumptions is not just relevant in the NHST context, but are also relevant in alternative analytic settings, such as Non-parametric or Bayesian statistical tests (e.g., (Zimmerman 1998; Gelman & Shalizi 2015)).

The statistical validity of analysis depends on a large set of factors, such as power analysis, experiment design, and representativeness of the sample taken. One very prominent factor is the need to choose the appropriate method for a given analysis setting. Often, data and experimental properties require a specific approach to be used. For example, to assess a mean difference between two conditions using a t-test (a commonly used *statistical method*), one first needs to ensure that the data (either directly for small samples or asymptotically) follow a normal distribution (Open Science Collaboration 2015; Fiske 2016). Depending on the degree of the violation, the validity of the results may be undermined, which may reduce the reproducibility or generalizability of the research (e.g., Glass et al. 2012). Hence, some analyses can be correct, even if the assumptions are violated, but this violation needs to be discussed and justified.

Thus, even if some methods are robust to violations of their assumptions in certain cases, it is essential to *report on these violations* to assess their impact on the analysis. Lack of reporting of assumptions in an article could mean that some assumptions were never tested, which therefore raises questions about the validity of the results attained *via* these statistical methods. A lack of reporting could also mean, that the statistical assumptions in an article were tested but are not described. Both conditions contribute to our *status-quo*, where reviewers are asked to almost blindly trust authors on their data analysis – a culture that stands in contrast to the scientific method.

A focus on assumption reporting naturally concerns itself with the products of the scientific research process (published papers). As we will show, the process currently does not emphasize assumption reporting. Assumptions are not emphasized in reviewing and authoring standards. The lack of attention to the issue of assumption reporting is perhaps further influenced by limited reviewer time and lack of a consistent, easy way of measuring whether statistical assumptions are reported.

This paper outlines a solution to these problems. Our contributions, in order of presentation, include:

- A *review of standards on reporting statistical assumptions* (along with a discussion of common statistical flaws and related issues). This includes both a literature review and an analysis of the materials provided to reviewers by journals in our sample.
- A *scalable method testing whether publications meet basic assumption reporting standards*. For this, the method relies on an extensible rule-base to determine the assumptions expected to be reported for the applied statistical methods and crowdsourcing to ascertain the presence of the assumption reporting in a paper. Our approach yields an F1-Score of 83%. We used this approach to analyze over 600 papers in the analyses described next.
- A *comparison of statistical assumption reporting of CHI with five top journals from different fields* (medicine, psychology and management). To do so, we empirically determined the most frequently used statistical methods in these journals (i.e., ANOVA, linear regression, logistic regression, Chi-square test, and t-test). Our results indicate that across the disciplines, on average, three out of four papers lack any reporting of the statistical assumptions or a discussion thereof. In addition, we find that regardless of the field, only very

few assumptions (less than 6%) are reported for these frequently used methods.

- An *analysis of the reporting of statistical assumptions at CHI over time*, where we sampled a total of 261 papers over 25 years (from 1989 and 2016). Our data indicates that from the papers at CHI that use either the t-test or ANOVA, *most* (87%) *do not report any assumptions*. Additionally, we find evidence for only very mild improvement of the fraction of papers reporting at least one assumption over time.

Supported by these findings, *we argue that a discussion of how to develop a common ground on assumption reporting is warranted*.

2 Related work

Recently, the scientific community has expressed great concern with the quality of published research (Nosek et al. 2012; Mann 2016; Eicken 2013). This concern seems to be present across various disciplines, including social psychology, economics, or medicine. The discussion on this topic has intensified as a group of researchers, part of an Open Science Collaboration initiative, conducted a collective effort to replicate 100 experiments reported in papers published in three top-tier psychology journals during 2008 (Open Science Collaboration 2015). The authors were seeking to compare the results obtained in the original experiments with the reproduced experiments and assess whether the results are statistically coherent. They found that only about one-third to one-half of the original findings were also observed in the replication study. An underlying reason for the results of the Open Science Collaboration initiative is outlined by Nosek et al. (2012), who point to a conflict of interest embedded in the scientific practice for scientists: on one hand there is a desire to get published and on the other hand, there is a need for rigorous, clear, and accurate results. Recent work has proposed strategies and methods by which researchers can avoid biases and personal narratives inherited in their work, such as confirmation or publication bias (Jussim et al. 2015). In addition, rigorous experimental design is critical to scientific validity (Fiske 2016).

2.1 Correcting the use of statistical methods

However, statistical flaws remain one of the main factors hampering studies' validity and reproducibility. A large proportion of published studies contain at least statistical inconsistencies or, at worst, statistical errors (Strasak et al. 2007). Even though the taxonomy of such errors has not yet been established, they can roughly be classified into problems related to sample size and power analysis, little reporting, significance testing, causal inference and confounding, and last but not least, method selection and testing for assumptions (Rouder et al. 2016; Westfall & Yarkoni 2016; Strasak et al. 2007). For example, the selection of the sample size has to account for possible type II errors and can therefore be improved by power analysis. Moreover, the sample size should be always (pre-)calculated and fixed before the experimental stage and withdrawals during the experiment or data analysis have to be recorded and reported (Kuzon et al. 1997). Another prominent example is multiple hypotheses testing: when performing multiple significance tests (with the increasing risk of false positives), it is necessary to control for the familywise error rate (FWER), such that the overall error rate of the analysis is preserved.

One of the most common, and at the same time probably most criticized statistical concepts – p-value – has drawn special attention in the recent years (Nuzzo 2014). Criticism of the use of p-values has led some journals to the drastic step of banning the use of p-values as statistical measure altogether, because "*p-values are slippery, and sometimes, significant P-values vanish when experiments and statistical analyses are repeated*" (Woolston 2015). While p-values were initially intended to be used as evidence for a chance that there might be a difference between the researched groups worthy of a second look, they have turned out to be a mainstream statistical tool, inherently prone to misinterpretations and abuse. A well-known abuse is a phenomenon known as "P hacking" where authors try different hypotheses until a significant result is reached (Baker 2016b; Leek & Peng 2015).

The criticism of typical use and abuse of NHST raised in different communities such as economics and psychology has sparked similar discussions in other fields. For example, in the CHI community Kaptein and Robertson (2012) wonder if p-value based inference should continue to be practiced. Disillusion with traditional p-value based inference led researchers in CHI community to propose alternative approaches better fitting the experimental nature of the field. Kay *et al*. (2016) advocate for adopting Bayesian analysis as it is more precise with small sample studies, enables a better comparison of proposed novelties with current state, and, just as importantly, brings the shift from the "does it work?" question (i.e., rejecting the null hypothesis) to understanding of the effect size, and, therefore, leads to the core question of whether the effect size worth of attention and further research. Another alternative to NHST is Magnitude-based inference, which uses the smallest important effect in making an inference and, as a result, the effect is not a consequence of sample size which makes this methodology very useful in user research (Van Schaik & Weston 2016).

While experimental design and reporting are one prominent reason for flaws in published research, the use of incorrect statistical methods, given the characteristics of the data, and the goal of the analysis, is another. One of the main reasons for assumption checking is to help with selecting the right method of analysis (Strasak et al. 2007). A prominent example is the assumption of heteroscedasticity (i.e., the property that the variance of a variable across the range of values of a second variable that predicts it is about the same). This is important for multiple linear regression because a violation of this assumption can lead to the substantial distortion of findings and increase the chance of a Type I error (Strasak et al. 2007). Another example is the assumption of a linear relationship between independent and dependent variables in linear regression, which is necessary to adequately estimate the true relationship between variables. Violation of this assumption can lead to increased Type I and II errors and undermine the analysis' results (Osborne & Waters 2002). Sometimes an assumption can be overlooked due to other statistical aspects of data analysis. For example, the assumption of normality in a t-test can be considered unnecessary when the sample size is big enough (due to the central limit theorem). Very few, though, will argue that this consideration should be ignored altogether, especially since only rough rules of thumb are available to decide which sample sizes can be considered high enough for asymptotics to work. Instead, it should be better reported and and discussed.

As a final example, the interaction between assumptions can have an impact on method validity. For instance, the type I error rate of the t-test for the Pearson

product-moment correlation coefficient has been shown to tolerate violations of the normality assumption only under certain scenarios: When two variables are simulated to be fully independent, violations of normality do not affect the type I error rate at a nominal alpha level of 5%, whereas under alternative null hypothesis scenarios werethe two variables are simulated to show a correlation of zero but are not independent, the type I error rate is strongly affected by a lack of normality (Edgell & Noon 1984). Therefore, even though some statistical tests are somewhat robust towards violations of assumptions, it is still important to assess and discuss the assumptions, to ensure statistical validity.

2.2 Correcting the reporting of statistical methods

While correct use of statistical methods is critical for the scientific process and improvement of the validity of published works, correct reporting of statistical methods is equally important to the scientific process. It is through correct reporting that researchers can demonstrate correct application of the scientific process, including power analyses, sample selection, and testing of assumptions, which all can impact the meaningfulness and valid results (Affairs 1999; Good & Hardin 2012).

Fernandes-Taylor *et al.* (2011) surveyed editors and statistical reviewers of 20 high-impact medical journals and found that submissions fall short of adequate reporting completeness and suffer from statistical and sampling issues. Among other flaws, the respondents pointed out flaws in data analysis, such as violations of model assumptions and analysis errors, overlooking clusters in the data, and improperly addressing missing data. These concerns are part of a larger ongoing debate on the reproducibility of the research.

Additionally, standards differ by field. In medicine, guidelines published by biostatisticians encourage authors to mention whether their assumptions were met either in paper or in supplementary materials (Lang & Altman 2013; Smith 1990). In psychology, assumptions are often either not tested, not reported, or ignored (Nimon 2012; Erceg-Hurn & Mirosevich 2008). This may be due to a self-selection process, where during the review process authors convince reviewers that the assumptions are met or may be relieved. More likely, however, is a lack of awareness for the necessity to report on assumptions (Hoekstra et al. 2012). Another explanation might be skepticism as to the practicality of the existing methods to test assumptions besides visual inspection (e.g., Affairs 1999). This could have led to a culture where it is expected from authors to visually inspect the data, without a need to report on it in the paper.

To summarize related work so far, there is a large and ongoing discussion about a wide variety of factors impacting the validity of published scientific works. While much of this discussion focuses on the scientific production, we argue that the *reporting of scientific results* is as important as the *production of scientific results.* Regardless of what statistical analysis methods are used, one specific aspect of reporting that relates to the correct application of statistics in the scientific process is the reporting of checks on the underlying assumptions of statistical methods. Such assumptions are inherent to almost all statistical approaches (including alternatives to NHST such as Bayesian analysis). However, in this paper we will focus on

assumptions for NHST because it is still in wide use and it is the only set of statistical techniques for which there is a large amount of data that can be analyzed to assess the rigor of reporting of scientific results.

3. Crowdstat : A Tool for the analysis of assumption REPORTING in published works

While the focus of this paper is on assessing reporting rigor across multiple fields that use statistical analysis *at scale* (this paper assesses over 600 papers), an equally important contribution of our work is the creation of an automated method for assessment called CrowdStat. CrowdStat is based on the crowdsourcing and structured in a way that enables non-experts to take part in the evaluation of statistical assumptions reporting.

CrowdStat is useful not only for scalable assessment across fields, but also offers the possibility for authors and/or reviewers to quickly and inexpensively check whether the reporting of statistical assumptions in an article requires further investigation. Here we describe CrowdStat and present an analysis of its validity in identifying potentially problematic papers. While CrowdStat should not be used without human oversight, we believe that it has the potential to be a useful sort of first stage test for authors or reviewers.

CrowdStat has a three-phase process for determining whether assumptions are reported.[9] First, it automatically finds and pairs statistical methods and possible reports of the assumptions underlying them, eliminating any pairs which do not make sense (because an assumption is not at issue for a given method). Next, it creates images that show snippets of text around each candidate method/assumption pair. Finally, a human is asked to verify, using the text snippets, that the method and assumption truly are related.

CrowdStat was implemented in Scala language, and uses Apache PDFBox[10] to access a paper's text.

3.1 Automated Method-Assumption (*ma*) Extraction

CrowdStat uses a bag of words approach to find sentences that may refer to methods or assumptions. For a given paper, a viable pairing of method words and assumption words are considered to be a candidate method-assumption pair (denoted by *ma*). CrowdStat's rule base is used to select candidate *ma* pairs that are viable. For these valid *ma* pairs, image-snippets are created. The snippet contains the text of the paper between the occurring method and the assumption of the *ma* (see **Error! Reference source not found.** for an example).

[9] The code, the analysis, and data are available under an open source license on GitHub at https://github.com/pdeboer/PaperValidator

[10] https://pdfbox.apache.org/

> To evaluate **H1**, we tested in Study 1 (S1) if there were differences regarding the number of unknown words among CI^2 translations and its 3 base languages. We carried out a one-way (continuous factor) analysis of variance (ANOVA), since we verified that both normality and homocedasticity did hold between groups. Interaction effects were considered at the $p < .05$ level for each of the tested conditions. On the other hand, **H2** was evaluated on the basis of the following criteria:

Figure 2: Example of a snippet for a method-assumption pair ***ma***. Methods are highlighted in yellow, assumptions are highlighted in green. In this example, a test of the assumption of *normality* has been reported for the method *ANOVA*.

CrowdStat iteratively considers each statistical method and assumption pair, which are drawn from a pluggable rule base, σ, which corresponds to Table 1 in our studies. Each method and assumption may have synonyms (such as *normal distribution, bell curve,* and *normality* for the test of normality, or *homoscedasticity* and *equal variance* for correlation). For each synonym, CrowdStat constructs a regular expression that accounts for white space, page-breaks and hyphenation. All assumptions matched in the text for all methods are considered tuples of candidate method-assumption pairs *ma*.

Table 1: Methods and assumptions we used for our validation. Mapping according to (Field 2013)

	Normality	Homo-scedasticity	Linearity	20% rule	Independence	No Multi-collinearity
t-test	x	x			x	
ANOVA	x	x			x	
Linear regression	x	x	x		x	x
Logistic regression					x	
Chi Square				x	x	

Note that there are multiple possible candidate matches for each assumption, since some assumptions are shared by different statistical methods. For example, a *t-test* as well as *MANOVA* both require the underlying data in each group to be normally distributed. An author might report only one normality test and that might be correct (if both statistical tests are done on the same data). In addition, some synonyms of assumptions may match with words used in general English, e.g. *normal for normality*. Thus, CrowdStat next extracts snippets of text that can be shown to crowdworkers to assess the validity of the *ma*-pair.

CrowdStat generates a snippet containing the portion of the PDF where both relevant terms (method m and assumption a) of an ma-pair are highlighted in different colors (see Figure 2 for an example). For copyright reasons, snippets are cropped to .6 inches of text (about 3 lines) above and below the highlighted method and assumption. If the relevant method and assumption are more than one page apart, CrowdStat excludes all intermediate pages, again for copyright reasons.

3.2 Human Judgment of Relevance of *ma*-pairs

Given a candidate *ma* pair, CrowdStat next needs to check if it is a valid case of assumption *a* being reported for method *m*. More specifically, each method-assumption pair *ma* has to be checked, such that:

- The assumption candidate *a* is indeed a statistical assumption (e.g. "normal" is a commonly used word in English)
- The assumption candidate *a* is semantically relevant to the statistical method *m* under inspection.
- The authors need to report on some condition about the assumption (not necessarily that it passed). We were fairly lenient here, for example the following case would pass despite not giving any details about testing the assumption: "*even though the data was not normally distributed, we executed a t-test*")

Because authors often report on assumptions close to where statistical tests are mentioned, CrowdStat sorts all candidate assumptions for a given method by distance (defined as the number of characters between *m* and *a*) and sequentially asks crowd workers for their judgments. As soon as a synonym for *a* is positively associated with *m*, all other synonyms for *a* are discarded, for that method (since assumption *a* is now recorded as checked). This reduces the number of redundant checks and the overall cost of running the system.

For each *ma* pair, Crowd workers are shown a brief explanation of the task, stating that statistical methods have 'prerequisites' (assumptions), where the color used to highlight methods (yellow, in Figure 2) and their prerequisites (green, in Figure 2) was illustrated. Next, CrowdStat asks '*in the text above, is there any relationship between the prerequisite and the method*'? This question is accompanied by a hint, giving examples of direct relationships (such as '…we tested [PREREQUISITE] before we used [METHOD]…') and an indirect relationship (such as '…our data were tested for [PREREQUISITE]. Using [METHOD]…'). If the worker answers with "*Yes*," then CrowdStat asks '*Did the authors of the text confirm that they have thought about the prerequisite before applying the method.*' Additionally, CrowdStat asks the crowd workers to rate their confidence in their own answer on a Likert scale from 1-7, and for a brief text explanation of why they selected the answers they chose.

For increased reliability, CrowdStat employs a variable number of crowd workers for every *ma*, depending on crowd-disagreement and aggregate their results. More specifically, CrowdStat uses the crowdsourcing library PPLib with Beat-By-K and K=3 to determine the number of crowd workers required based on crowd disagreement during runtime (see de Boer & Bernstein, 2015) for more details on PPLib or Beat-By-K). Beat-By-K can be seen as an extension to a Majority Vote, which continues asking more crowd workers to submit votes until the most popular item has at least K more votes than the 2nd most popular item, where crowd workers can not submit more than one answer per snippet. In addition, CrowdStat only uses answers with confidence above or equal 5. This threshold was identified in pilot experiments. A higher threshold would likely increase the accuracy of CrowdStat, but lead to higher expenses.

Through this process, CrowdStat collects a list of all reported *ma* pairs. It then automatically generates a list of *non-reported* (but expected) *ma* pairs using $\boldsymbol{\sigma}$. Thus, the final output of CrowdStat is a list of assumptions that are not reported (but should have been) in the candidate paper.

3.3 Validation of Crowdstat

Prior to applying CrowdStat at scale, we performed a validation of our approach. To do so, we applied our system to 50 papers sampled from the CHI conference and compared the results with manually annotated papers.

4.Method Overview

To validate CrowdStat, we empirically determined the two most used statistical methods in CHI. We obtained the papers published in CHI (full papers and notes, no posters, keynotes, *etc.*) between 1989-2016 (5132 in total). We constructed a dictionary of statistical terms (based on the article names in Wikipedia's 'statistics' category). We found that the use of statistics in CHI follows a long-tail distribution, where few papers use many statistical terms and the vast majority use very few of them (see Figure 3). We selected the 38 most common statistical methods and calculated the fraction of papers per year using each method. Out of these 38 most common methods, *ANOVA* is the most common test (27% of papers use it), followed by *t-test* (8%) and *Linear regression* (3%). To limit cost, our analysis therefore focused on the 3 methods used most over the years: *ANOVA* (and its derivatives *MANOVA* and *ANCOVA*), *t-test* and *linear regression*. This left 1671 papers.

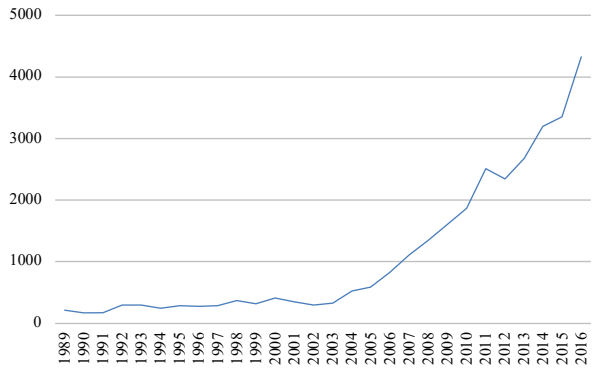

Figure 3: Statistical terms used in CHI

4.1 Sample

In order to be able to reliably estimate the accuracy of our approach, we used the following formula to derive the necessary sample size

$$n \geq \left(\frac{z}{m}\right)^2 \cdot \hat{p}(1-\hat{p}) = 191 \geq \left(\frac{1.64}{0.05}\right)^2 \cdot 0.77(1-0.77) \qquad (1)$$

This formula is designed for constructing a confidence interval for a proportion with 90% confidence and an error margin of 5%, so that z is the quantile of the cumulative normal distribution at 90% confidence (=1.64), m our targeted error margin of 5% and p is the estimated proportion, i.e. the accuracy of our approach. With no information about p, we would calculate the sample size with a value of p =0.5 which would lead to the highest possible sample size with all other parameters fixed.

However, the minimum accuracy calculated across all of our pilot experiments was 77%, which we conservatively used as p. The resulting n is therefore 191 ma pairs. In preliminary experiments, we observed that papers had an average of two ma pairs. Therefore, we had to sample a total of $191/2 < 96$ papers to meet our power requirements.
To be on the safe side, we rounded the result up to 100 papers and sampled equally across the two categories of papers using many and few statistical terms as follows: We sorted papers by number of method words used in the paper, and selected the 50 papers using the most statistical terms and the 50 papers using the fewest statistical terms for our sample. Our goal in sampling from both ends of the spectrum was to increase the likelihood that we had included both good and bad examples in our sample.

4.2 Ground Truth Data Collection

Three of the authors with advanced statistics training manually annotated 100 of the papers. Of these, 11 papers were annotated by all annotators so we could assess the inter-rater agreement. Annotators were given a PDF for each paper and a list with methods and corresponding assumptions (see **Error! Reference source not found.**), w hich were also highlighted in the PDF. Annotators recorded all *ma* pairs occurring in a paper as *reported*. Any missing *ma* pairs were recorded as *not reported.* For papers where at least one assumption was reported, we calculated the inter-rater agreement using Fleiss' Kappa, which was 0.79. This is conventionally interpreted as *substantial agreement*. Kappa is much higher if we also include all papers with no assumption reporting, as raters are most likely to agree on the absence of any relevant assumptions.[11]

Among the 100 papers analyzed, only 11 included at least one *ma* pair. Since only very few papers reported on any assumptions at all, all papers reporting at least one assumption were later given to all annotators to increase labeling validity. As a baseline for our system, we then used the result of a majority vote among the three candidate annotations for each *ma* pair.

At the end of this process, annotators agreed on 82 *reported ma pairs* among the 11 papers. In addition, 59 expected *ma* pairs were absent from the 11 papers. If all assumptions had been reported, across all 100 papers, there should have been 250 *ma* pairs.

4.3 CrowdStat Data Collection

We next ran CrowdStat on the 100 papers that had been labeled, using Amazon Mechanical Turk (MTurk). For each paper, for each possible *ma* pair, CrowdStat recorded whether it was *reported* or *not reported*. In addition, as described earlier CrowdStat collected a (free-text) justification with each submission and rater confidence.

[11] Given that there are much more papers without any *ma* pairs we have a highly-unbalanced distribution of papers containing and papers not containing *ma* pairs. Hence, a Kappa including these papers would be unreasonably high and, in our opinion, somewhat misleading.

We hired only US workers with less than 4% rejected HITs and more than 4000 approved HITs. Paper order was randomized, and HITs were posted on mornings of working days in the U.S. Eastern Daylight Time zone. Each MTurk task was priced at 40 cents and for each *ma* pair a crowd worker could only give one answer. The total cost of running the system on these 50+50 papers was $214 USD. Upon completion, we excluded one paper from the analysis, due to an extraordinarily high error rate in our tool. This paper discussed statistical methods at a meta-level (*e.g.*, presenting a tool for helping people do statistics) rather than applying these methods. After excluding this paper; 198 *ma* pairs remained, which still provides sufficient power (n>191).

At the end of this process, CrowdStat identified 83 *ma* pairs that were *reported* and 115 *ma* pairs that were *not reported*. Table 2 compares CrowdStat to the ground truth data. Of the 83 reported *ma* pairs, 68 were correct and 15 were false positives. In addition, there were 14 were false negatives among the 115 not reported *ma* pairs, an accuracy of 85%.

Table 2: Confusion matrix of CrowdStat when compared to manual annotation. CrowdStat correctly identified 68 *reported* ***ma*** pairs, and only 15 ***ma*** false positives and 14 false negatives.

		CrowdStat (Predicted)	
		ma reported	*ma* not reported
Ground Truth	*ma* reported	68 (TP)	14 (FN)
	ma not reported	15 (FP)	101 (TN)

4.4 Results and Discussion

To assess performance, we calculated *Precision*, *Recall* and *Cohen's Kappa*, using *reported* as the positive label. CrowdStat achieved a precision of 82% with a recall of 83%. The F1-score was also 83%. Cohen's Kappa was 0.70; CrowdStat performed slightly worse than expert labelers (who had an inter-rater agreement of .79).

We analyzed the 500 crowdworker comments given for the 29 false positive and false negative *ma*-pairs (recall that each *ma*-pair was rated by a variable number of crowdworkers). In our analysis, we paid attention to non-verbal cues indicating rushing (such as frequent misspelling) and cheating (such as very rapid response time or answers that had obviously been copied and pasted between questions) as well as the text content of the comments themselves. We found three common reasons for errors: (i) crowd workers who apparently rushed or cheated, (ii) crowd workers who did not understand what was meant by *method* and *assumption*, and (iii) crowd workers who misunderstand the task goal. For example, some workers analyzed sentence structure instead of the semantic relationship between the two marked terms in a snippet.

CrowdStat's high performance suggests that it is a viable means to automatically evaluate statistical assumption reporting in submitted papers. An automated approach for testing whether publications meet basic assumption reporting standards would (1) make it easier to track reporting quality over time, (2) provide authors with a simple way to check whether they are meeting a common standard, and (3) provide editors with a simple way to flag potentially problematic papers for further investigation. Note that we are by no means proposing to reject papers on

this basis, simply to provide some automated support for encoding and raising awareness about reporting standards.

Before a system like CrowdStat is used at a conference like CHI, the community needs to decide whether reporting of statistical assumptions is important. This small preliminary study suggests that until now reporting has not been considered as an agreed-upon standard. In fact, it may be surprisingly rare, as only 11 of the 100 papers we sampled reported *any* assumption, and only 50% of assumptions were reported (for papers that report assumptions at all). As we will show later in the next section, these numbers are not simply a matter of sampling error due to our small sample, but representative of not only the field of Human-Computer-Interaction but also many other fields.

5. Comparing reporting in different fields

Motivated by these results, we conducted a larger-scale analysis of 1-3 top journals from three additional fields. Note that we do not claim to be able to infer the state of a whole field by analyzing a single publication venue. Rather, we see the standards on statistical assumptions reporting in these journals as an *indicator* of the current expectations in these disciplines.

Our analysis focused on four fields: Human-computer interaction (HCI), management, medicine and psychology. These fields were selected because they all make frequent use of statistics. We chose psychology because of its use of statistical methods similar to those used in HCI. We chose medicine because of its long history of rigorous evaluation. In fact, modern statistics was often further developed as a remedy to a shortage in tools for rigorous evaluation in medical studies (e.g. Martin Bland & Altman, 1986). Finally, we chose management to further increase the range of our sample and thus its generalizability.

Note that we do not aim to prove statistical significance during this analysis, which would require a much bigger sample size drawn from many more journals and is beyond the scope of this research due to high costs and limitations imposed by publishers.

5.1 Journals

Journals were selected based on the following criteria: We selected (i) the highest rated journal (as ranked by ThomsonReuters InCites Impact Factor (IF) in 2015) that (ii) commonly included papers making use of statistical methods and (iii) whose publisher agreed to give us a license to apply text mining techniques to the sampled papers. Our sample focused on the year 2014. The six journals selected for our analysis are the following:

- BMC Medicine (IF=8.01): an open access, open peer-reviewed general medical journal, publishing research in all areas of clinical practice, translational medicine, public health, policy, and general topics of interest to the biomedical research community. During 2014, the journal has published 588 papers.
- *BMJ open* (IF=2.56): is a peer-reviewed open access medical journal, publishing research across all medical disciplines and therapeutic areas. As the journal is focusing on research relevant to patients and clinicians, it does not publish studies conducted in animals or laboratory studies not linked to patient

outcomes. Besides case-studies, the journal accepts all types of research, including but not limited to study protocols, phase I trials and meta-analyses. During 2014, the journal published 944 papers.

- *New England Journal of Medicine* (IF=59.55): is among the most prestigious peer-reviewed medical journals, with its first issue tracing back to 1812. The journal usually has the highest impact factor of the journals of internal and general medicine. During the last 26 years, the journal provided (delayed) free access to published studies six months after the publishing date. In 2014, the New England Journal of Medicine published 1075 studies.
- *CHI proceedings:* ACM conference on human factors in computing systems is a flagship conference in the field of human-computer interaction and one of the most prestigious in computer science tracing back to 1982. Papers published in CHI go through double-blind peer review process with an acceptance rate of about 23%. In 2014, a 449 papers were published.
- *Cognitive Psychology Journal* (IF=4.53): is a peer reviewed journal, publishing studies about attention, memory, language processing, perception, problem solving, and thinking. The journal was established in 1970 by Elsevier and publishes eight issues per year. The Cognitive Psychology journal had 34 papers published in 2014.
- *Journal of Management* (IF=6.05): is a peer-reviewed journal in the field of management published bi-monthly and dedicated to empirical and theoretical research. The journal covers domains such as business strategy and policy, entrepreneurship, human resource management, organizational behavior, and organizational theory. Seventy-three articles were published during 2014.

Before proceeding with our analysis, we reviewed the submission guidelines of papers, to better understand the expectations from authors with regards to reporting statistical assumptions. Surprisingly, many of the papers published in NEJM do not mention any assumptions. Reviewing the guidelines for the NEJM, we could not find any concrete guidelines[12] with regards to the assumption reporting. Interestingly, the guidelines do address the importance of the normality assumption when using the t-test, but do not provide any further expectations as to assumptions in general. In the BMC guidelines authors are asked to justify the appropriateness of the statistical test used. Specifically, authors are referred to the SAMPL (Statistical Analyses and Methods in the Published Literature) guidelines (Lang & Altman 2013) where they are tasked to "*verify that data conformed to assumptions of the test used to analyze them*". In addition, authors are especially asked to pay attention whether the data is skewed, that paired data is analyzed with paired tests and when applying linear regression, that the underlying relationship is indeed linear. This means that only some assumptions are highlighted for the reviewers. BMJ has a very detailed checklist to follow for different types of studies (observational, meta-analysis, *etc.*). While the checklist[13] addresses various aspects of statistic validity, such as potential biases and data sources, it only provides a general request to describe all statistical

[12] http://www.nejm.org/page/author-center/manuscript-submission

[13] http://bmjopen.bmj.com/pages/authors/

methods used (e.g. STROBE[14] checklist). The Cognitive Psychology journal also does not specify whether the assumptions should be reported, but rather encourage authors to provide sufficient details to allow reproduction of their work.[15] Lastly, CHI[16] allows to submit supplementary material, but no instructions on statistical reporting are provided. We could not find any instructions of Journal of Management[17] regarding statistical assumptions either.

5.2 Sample

We obtained all published papers of the selected journals for 2014 as PDFs. We compiled a list of 30 statistical methods based on our earlier dictionary and a review of statistical literature (see Kanji 2006; Sheskin 2004 and Appendix I). We then submitted this list of methods to two professional statisticians to validate that the entries are indeed a reasonably common statistical methods and to assure that methods selected had at least one assumption that ought to be reported. We counted the occurrences of these methods in the sample and selected the five most common (based on number of occurrences in the sample) for further analysis: *t-test, ANOVA, Linear regression, Logistic regression and Chi Square test*. For these methods, we conducted a literature review and consulted with additional statisticians in order to create the final list of assumptions for each method, as well as the appropriate synonyms, that were then used to build the rule base for CrowdStat.

Of course, this final list had to some degree be a compromise. For example, some sources included the appropriate type of response variable for certain methods in their list of assumptions (such as a binary response variable for logistic regression). However, since this may be considered too trivial to report (and most statistical software packages would not even produce a result if a wrong type of response variable was used in the analysis), we did not include this in our final list of assumptions. As a general rule to avoid excessively penalizing papers, we included in our list of assumptions only those for which there was general consensus about relevance. One assumption that we could only include in a simplified manner, on the other hand, is the normality assumption, for example for the t-test. While normality is a necessary assumption here, that should be tested in smaller samples, it is asymptotically given if the sample size is big enough. Therefore, it could be that some researchers, who are used to analyze small sample sizes, would report this assumption, while others, who are usually handling big sample sizes, might not bother reporting on this assumption considering it unnecessary. As argued earlier, however, even if an assumption need not be formally tested for a good reason, we consider it important to report whether and why it holds. The finally used methods with their corresponding assumptions and synonyms can be found in Table 3Table 5.

[14] https://strobe-statement.org/index.php?id=available-checklists

[15] https://www.elsevier.com/journals/cognitive-psychology/0010-0285/guide-for-authors

[16] https://chi2017.acm.org/papers.html

[17] https://us.sagepub.com/en-us/nam/journal-of-management/journal201724#submission-guidelines

About ten percent of the total papers published in the three medical journals, and over thirty percent of papers in Journal of Management and CHI proceedings used the statistical methods we selected (most of the remainder did not make use of any chosen statistical method and, hence, given the long tailed distribution of method, most of these made no use of statistical methods). Our goal was to achieve a comparable magnitude to our pilot study, namely 191 *ma* pairs per journal. Thus, we randomly sampled papers of a given journal which used the selected statistical methods until we reached 191 *ma* pairs. Given that we could not find 191 *ma* pairs for CHI and Cognitive psychology in their 2014 volume, we included all papers that included such pairs resulting in 185 and 115 *ma* pairs for those publication outlets. This led to a total of 442 papers (as some of these might contain multiple *ma* pairs) across all journals (see also Table 3). The lowest number of papers sampled was 26 (Cognitive Psychology) and the highest was 139 (CHI). For cost-saving reasons, a professional statistician reviewed papers containing more than 30 *ma* pairs (92 papers), leaving 350 papers for analysis with CrowdStat.

Table 3: Number of papers using different statistical methods in every journal. In parenthesis is the percentage of papers using each method, out of papers sampled from every outlet (All numbers represent papers published during 2014). Techniques used in more than half of papers are sh.

Journal (Total sampled)	ANOVA	Linear Regression	Logistic Regression	Chi Squared	t-test	# *ma* pairs	# qualified papers (>=1 *m*)	# correct papers (>=1 *a*)	# papers published
BMC (TOT)	19 (29%)	39 (59%)	27 (41%)	38 (58%)	20 (30%)	316	66	42	588
BMJ (TOT)	31 (32%)	56 (58%)	44 (45%)	69 (71%)	38 (39%)	324	97	48	944
NEJM (TOT)	16 (19%)	35 (41%)	37 (43%)	38 (44%)	18 (21%)	282	86	55	1075
CHI (TOT)	77 (55%)	36 (26%)	20 (14%)	37 (27%)	62 (45%)	185	139	30	449
Cog Psych. (TOT)	19 (73%)	10 (39%)	10 (39%)	6 (23%)	8 (31%)	115	26	14	34
J. Manage.. (TOT)	10 (36%)	18 (64%)	5 (18%)	21 (75%)	9 (32%)	216	28	21	73
Mean %age	28.66 (41%)	32.33 (48%)	23.83 (33%)	34.83 (50%)	25.83 (33%)	239.67	73.67	35	

5.3 Results

In this was exploratory work, we did not begin this analysis with specific hypothesis in mind. However, we focus our reporting on the following questions:

1. Do fields differ in which methods they use (Table 3)
2. Do fields differ in which assumptions they report (Table 4)
3. Do fields differ in the extent to which assumptions are reported (Table 5)

The medical journals we analyzed (BMC, BMJ and NEJM) use Chi Squared tests and linear regression most frequently with the average percentage being 58% and 53%, respectively (see Table 3). The Cognitive Psychology journal uses mostly ANOVA (73%), while papers published in Journal of Management use Chi Squared tests (75%) and Linear regression (64%) much more than any other methods. These results

suggest that different disciplines have different preferences in statistical methods. The most commonly used methods across all fields are linear regression and Chi Squared tests, used 48% and 50% respectively. All journals use t-tests with roughly equal frequency (μ=33% of sampled papers).

Table 4: **Ratio of** papers where **no assumption** of the corresponding method is mentioned in paper to papers **using the corresponding method**. Highest ratio for each journal is shown with a red background; lowest is shown with green background.

Journal	ANOVA	Linear Regression	Logistic Regression	Chi Sq. test	t-test	Journal mean
BMC	68% (13/19)	87% (34/39)	85% (23/27)	84% (32/38)	75% (15/20)	80%
BMJ	77% (24/31)	86% (48/56)	80% (35/44)	81% (56/69)	71% (27/38)	79%
NEJM	100% (16/16)	100% (35/35)	95% (35/37)	89% (34/38)	100% (18/18)	97%
CHI	79% (61/77)	83% (30/36)	65% (13/20)	84% (31/37)	81% (50/62)	78%
Cog. P.	79% (15/19)	60% (6/10)	70% (7/10)	83% (5/6)	50% (4/8)	68%
J. Man.	70% (7/10)	72% (13/18)	80% (4/5)	62% (13/21)	67% (6/9)	70%
Mean	79%	81%	79%	81%	74%	79%

The 442 papers of our analysis employed 869 methods and 336 uses of methods for which no assumptions were reported. For every method, Table 4 shows the ratio of papers that *report **no** assumption* to those that use the mentioned method. For example, of the 19 papers published in our BMC sample that use ANOVA, 13 papers do not mention any assumption for ANOVA. All methods have four out of five papers not mentioning any assumption across journals, with t-test having a slightly better reporting. Surprisingly, the NEJM has the highest percentage of methods with no assumptions reporting (97%), while BMC, BMJ, and CHI proceedings hover around 80%, whilst Cognitive Psychology and Journal of Management have a slightly better performance, with about 70% of methods lacking any mention of assumptions.

To understand whether the assumptions for statistical methods are outlined in the supplementary materials, we manually reviewed all publicly available files belonging to the 442 papers we analyzed. The Cognitive Psychology Journal has all supplementary materials appended to the published manuscripts. This means that while exist, it is part of the published manuscript and, therefore was retrieved and analyzed by CrowdStat. The same is true for the Journal of Management, where all the relevant information is part of the published paper. In the proceedings of CHI we could only find videos discussing the papers. However, these videos never discuss the assumptions of statistical methods but rather focus on a high level explanations of the study. Out of the 66 papers published in BMC journal, 36 had supplementary materials and only 2 papers discussed statistical assumptions of the methods used in their study. In our sample of 97 papers in the BMJ, 24 papers had supplementary material whereas only 3 papers mentioned relevant statistical assumptions. A for the NEJM, we found supplemented materials for 74 papers out of 86 papers. Often this was a study protocol summarizing the whole aspects of the

conducted clinical trials. While these protocols discuss in great details the course of study they do not discuss the statistical assumptions. Out of 74 papers we found only 7 papers mentioning statistical assumptions.

The ratio of methods with *no assumption* to methods with *any assumption* reported might be seen as a generous measure. A more stringent one is to examine the amount every assumption is reported (in comparison to the number of opportunities to report that assumption). We show the ratio of *ma pairs found* to those *expected*, for each method and venue, in Table 5.

Interestingly, all journals have similarly low assumption checking rates. In other words, if one would expect all assumptions to be reported for every method applied in an article, all journals in our analysis would not live up to this expectation (the reporting rate ranging between 3% and 14%). Examining the vantage point of statistical methods, the results look similarly abysmal. ANOVA's assumptions are checked only in 5% of cases, and for t-test, linear regression and Chi square it is just 2.7%.

Table 5: Assumption reporting rate for every method, assumption and journal (ratio of *ma* pairs found to those expected). Values above 20% are highlighted, values above 10% are shown in bold.

	Ass.	BMC	BMJ	CHI	Cognitive Psychology	Journal of Management	NEJM	Mean	Average per method
t test	Normality	0	5.3%	8.1%	12.5%	0	0	4.3%	2.7%
	Independence	0	2.6%	0	0	11.1%	0	2.3%	
	Homogeneity of variance	0	0	9.7%	0	0	0	1.6%	
ANOVA	Normality	15.8%	19.4%	9.1%	5.3%	0	0	8.3%	4.9%
	Independence	5.3%	0	0	0	20%	0	4.2%	
	Homogeneity of variance	0	3.2%	10.4%	0	0	0	2.3%	
Linear regression	Homogeneity Of variance	0	0	0	0	0	0	0	2.3%
	Normality	2.6%	3.6%	0	0	5.6%	0	2%	
	Lack of multi-colinearity	0	0	0	10%	5.6%	0	2.6%	
	Linearity	0	0	0	10%	11.1%	0	3.5%	
	Independence	0	3.6%	6.2%	0	11.1%	0	3.5%	
Chi square	Expected cell count	0	0	0	0	0	0	0	2.7%
	Independence	2.6%	0	2.7%	16.7%	4.8%	5.3%	5.4%	
Logistic regression	Independence	7.4%	11.4%	35%	20%	20%	5.4%	16.5%	16.5%
Weighted Average by number of papers		6.1%	9.4%	11.9%	13.7%	13.2%	3.5%		

5.4 Discussion of cross-field analysis

The low assumption reporting rates do disclose an ambivalent state of affairs. Interestingly, we could not identify any assumption that stands out in reporting rate. For instance, one might assume that normality is much less important due to the central limit theorem, and since many statistical methods are robust to its violation. This suggests that the issue is not a matter of certain assumption but rather a general lack of perceived importance of assumption reporting.

The results also suggest that there are different expectations for different methods. While reporting on the assumptions of some methods is more rigorously followed, other methods are either implicitly considered to be robust to assumption violation or overlooked by authors due to lack of knowledge. Interestingly, the standards between journals representing different disciplines substantially vary. For instance, even though linear regression is a method that is overall most often used without any assumptions mentioned, it is much more strictly reported in Journal of Management and Cognitive Psychology. On the other hand, ANOVA is rarely used without any assumption mentioned in medical journals (18% on average), while in CHI assumptions are not mentioned 60 percent of the time.

6. Analysis: assumption reporting in CHI over time

The rarity of statistical assumption reporting is similar in all of the fields we sampled. However, we were curious whether there is a change in the reporting of assumptions over time. Specifically, we wanted to test the following hypothesis:

Has the reporting of assumptions improved over time?

To assess this in a cost-effective manner, we focused on the CHI conference, for which over 20 years of papers were freely available to us.

6.1 Data Collected and Analysis Method

To better quantify this effect in CHI, we analyzed papers published between 1989 and 2016. For cost effectiveness, we focused our analysis on the two most popular methods (ANOVA and t-test). Additionally, we only sampled once every five years (specifically, the years 1989, 1994, 1999, 2004, 2009, 2014). From these years, we collected all papers that mentioned the two methods and then sampled 30 papers randomly from those papers. If fewer than 30 papers qualified, we used all qualified papers. This occurred three times (1989: 8 papers, 1994: 9 papers and 1999: 17 papers. We therefore increased our sample to additionally include years adjacent to the ones lacking magnitude: 1995: 3 papers, 1996: 14 papers, 2005, 2006, as well as the more recent 2015 and 2016 to see if there was a recent trend indicating change.

In total, our sample included 261 papers across all years under analysis. Note that our system did not find any *ma* pairs in any papers published in 1995 and 1999. We therefore manually checked each paper in our sample of these years, to ensure that the results were correct (they were). The total cost of running CrowdStat was $287 USD.

We calculated two metrics: **ShareReported is** the average share of assumptions reported, *per* year (the ratio of *ma* pairs found to *ma* pairs expected, per paper); **AnyReported is** the percentage of papers *per* year reporting *at least one* assumption.

Our primary analysis method was linear regression. A positive slope would mean an (approximately linear) improvement over time. Because of the uneven nature of our sample (fewer papers in earlier years), we also used a weighted linear regression, where we weighted the fractions of each year by the number of samples (i.e. papers) in that year. We visually inspected the QQ-plots to evaluate the normality as well as the residual plots to examine whether there are obvious deviations from mean zero among the residuals. The multicolinearity assumption is not relevant as we only have one explanatory variable – years. The QQ plot of **AnyReported** is deviating from the normal distribution with skewed distribution towards the recent years. Even though the plots do not conclusively prove the assumptions to be met (Osborne & Waters 2002), violations are fairly mild and may result from the small sample size (N = 12 years). The reported results should therefore be seen in light of these constraints.

Since the unweight data for both ShareReported and AnyReported violate the normality and homoscedasticity assumptions (and the assumptions are inconclusive in the weighted model), we also calculated ShareReported and AnyReported for two larger time windows 2008 or earlier, and 2009 or later. These years were selected because they approximately divided our sample in half. There were 141 papers in our sample from 2008 or earlier and 120 from 2009 or later.

We used a Mann-Whitney U test to compare ShareReported and AnyReported between the two groups. Mann-Whitney U is non-parametric and therefore does not assume normality. While originally proposed without requiring homoscedasticity, many statisticians currently argue about lower confidence in results in cases of unequal variance (Vargha & Delaney 1998). Zimmerman (Zimmerman 2004) quantifies this effect for various proportions between the standard deviations of the two groups. In our case, this proportion is 1.79.

To quantify the effect size, the Rank Biserial Correlation is often used for Mann-Whitney U. Another part of the same research question is, whether the aggregates have changed over time. **W**e therefore created a linear regression on the fraction of papers reporting *at least one* assumption for each year.

6.2 Results

Our results show a low percentage of papers reporting any assumption, with a very moderate increase. Only 35 of the 261 sampled papers report at least one of their assumptions. Figure 4 shows the share of papers checking at least one of their assumptions over time. A slight upward trend towards more papers reporting some of their assumptions is visible on the graph:

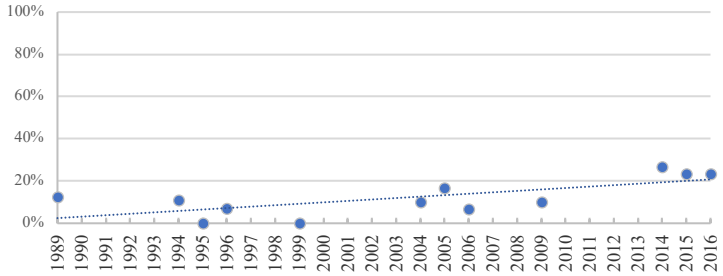

Figure 4: share of papers reporting at least one assumption (AnyReported) in CHI over the years

An unweight linear regression on the AnyReported metric (presented in Figure Figure 4) seems to point to a small increase in the share of papers reporting at least one assumption over time (slope=0.0066≈0.7% increase per year, SD=0.0022 with R^2_{adj}=0.4). When weighting each year by the sample size in that year, the effect is slightly increased, with a higher R^2_{adj}=0.53 (slope=0.0089≈0. 9% increase per year, SD=0.0024).

Among the papers reporting assumptions, roughly half of the expected assumptions are tested as shown in Figure 5. While there is no obvious trend, it seems that since 2004 the ratio is much more stable and is around 50% reporting rate.

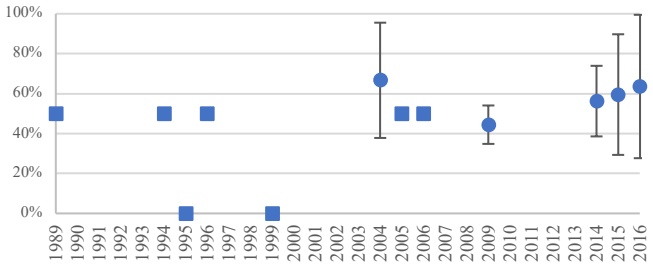

Figure 5: ShareReported (y axis) *vs* year (x axis). Mean percentage of assumptions reported in papers that check at least one assumption, (100% would mean that in that year, all assumptions were mentioned). The error bars (mean ± 1 standard deviation) are only shown for years with more than 3 papers reporting at least one assumption (circles).

The weighted regression on the ShareReported shows a small improvement over time, but only 13% of the variance can be explained (i.e. R^2_{adj}) with our linear model (slope=0.01≈1% increase per year, SD=0.0063). Recall that this analysis does not meet its statistical assumptions, which further reduces explanatory power. Hence, we also ran the Mann-Whitney-U test. This test shows a slight difference between assumption reporting prior to 2009 and afterwards (p=0.001). A Rank Biserial Correlation quantifies the effect size as very weak ($r_{RB} = 0.14$)

6.3 Preliminary Conclusions for Assumption Reporting in CHI over time

Since, only a small fraction (13%) of papers report checking assumptions, it is difficult to draw strong conclusions from these data (except that CHI may wish to reconsider its standard for what is expected with respect to statistical – and particularly statistical assumptions – reporting).

Whether or not the community is already making a change as issues with NHST receive increasing attention is unclear. However, our data indicates a near-linear trend developing towards a larger fraction of papers reporting at least one assumption (with a very low slope, meaning a slow rate of increase).

7. Discussion

Our analysis clearly indicates different preferences for statistical methods across journals associated with different disciplines. For example, while venues that focus on more behavioral methods such as Cognitive Psychology and CHI tend to use

ANOVA more, medical journals tend to use more Linear Regression and Chi Squared tests. Cultural predisposition might lead to a lock-in where researchers reuse methods they are well aware of and know how to operationalize properly. If a method is often used in a field, researchers might assume that the use of said method will not raise doubts among reviewers. Also, in some fields obtaining samples is very costly—or even impossible—and therefore researchers use statistical methods known to perform better with small sample size. There are a subset of methods known to be robust to small sample sizes and therefore researchers may limiting themselves to use these methods. Sometimes data has certain characteristics that constrain a choice of method. For instance, if data is time-dependent, skewed, or non-parametric, the method of choice will need to reflect these constraints.

The results of this study were surprising to us, as one would expected journals with higher Impact Factor to exhibit more rigorous standards. In practice, all analyzed journals showed very low assumption reporting rates. The low reporting rate of statistical assumptions is striking. Even journals with very high impact factor seem to not have clear standards on statistics reporting when it comes to assumptions.

Venues prioritize specific assumptions differently. For example, papers in our sample from the Journal of Management report relatively often on the independence of observations but never report on homogeneity of variance. On the other hand, papers in our sample from the New England Journal of Medicine do not report on any of the assumptions except independence of observations while using Chi Square or Logistic Regression.

All venues had low assumption reporting rates (if we consider all assumptions of every method used). When we use the more generous metric of mentioning **any** of a method's assumptions, there was a mild difference. While 70% of the Journal of Management papers in our sample do not mention any assumption, almost all (97%) of NEJM do not refer to any of the underlying assumptions. This might be interpreted by perceived unimportance *to account for all theoretical assumptions* since they are considered impractical and unnecessary. Alternatively, authors think that it is clear that they check assumptions and therefore not worthwhile reporting.

8. Limitations

The threats to validity can be classified into 2 categories: sample-related and approach-related. While the former is concerned with the statistical power of our analysis, the latter can have negative impact on the reliability.

8.1 Sample-related threats to validity

Our sample of at most 30 CHI papers per year may not be enough to draw a conclusion. This issue is aggravated in earlier years of CHI, where fewer papers used our statistical methods under inspection. This problem could be addressed in future work by including more statistical methods and hereby broaden the scope of the analysis.

Second, in similar vein, our analysis of CHI papers over the years is limited to the two most commonly occurring tests with assumptions – the t-test and ANOVA. It is possible, that CHI papers are better at reporting assumptions for other, less common

statistical methods. However, our focus on the two most prominent methods ensures that our insights apply to the majority of CHI papers using NHST.

Third, in hindsight, the assumption made for the power analysis of the *validation*-dataset, that every paper would have about two assumptions, was somewhat oversimplified. Ultimately, our power goal was still reached though for the CHI papers that have been used to validate our approach.

Fourth, our cross-field study analysis lacks sufficient power to draw inclusive conclusions. Due to the complexity and cost of analyzing an extensive corpus of papers we limited our study to few journals with high Impact Factors. In this study our goal was to gain a better understanding of the current state of other fields. Future research can address issues such as generalizability of journals to a field and increase the sample to allow for a more thorough statistical analysis.

8.2 Approach-related threats to validity

While the validation of our system covered both, step A (extraction of *ma*-pairs) and step B (filtering unreported *ma*-pairs), it is possible that the analysis of CHI over time would have required different sets of synonyms (*e.g.,* assumptions might have been addressed differently in earlier years unbeknownst to us). Or, more generally, that our rules for identifying assumption reporting may be too conservative (relevant for step A). Moreover, while performing step A (extraction of *ma* pairs), CrowdStat might generate much more potential *ma*'s for assumptions like *independence* and *normality* than for assumptions like *multicolinearity.* This is due to the natural prevalence of these words in English and since they are often used in different contexts in addition to the statistical assumptions. This might potentially lead to more false positives for these assumptions and result in overestimation of these assumptions reporting rate.

Second, the current version of our StatCheck does not differentiate between the meta-studies where a statistical method is mentioned but not applied (as in the two papers, which we had to exclude in our *validation*-dataset) and papers where a method is mentioned and used for analysis. Whilst we corrected for this error when testing the validity of our approach, we did not do so in our CHI analysis, which may add another source of error. We believe, however, that such an effect would be limited and equally distributed over the years. As mentioned, this could be fixed in a next version of our system.

Third, our crowd process may be unreliable, returning different results when being given the same inputs. While non-determinism in a crowd setting is well-known, a strong variance of the process would reduce its generalizability. However, our validation indicates that the variance is close to the variance of expert raters. We would need to repeat this testing for other domains to ensure the generalizability of our method.

Last, CrowdStat is imperfect, with a Precision and Recall 82% and 83% respectively of our system – implying that the results can err to some extent. However, even human annotators could be expected to make some errors as evidenced by the inter-rater agreement (Fleiss' Kappa) of .79 achieved by the authors when establishing a baseline (Cohen's Kappa of our system is .7). Since the reviewers of CHI or any other

scientific peer-review outlet do not necessarily always agree, our system's performance is close to the best available alternative.

9. Conclusions

Our results provide strong evidence that standards for reporting assumptions of statistical methods are insufficient across research fields. Additionally, the standards regarding which assumptions are more important and should be reported vary between fields. We argue that the low standards of statistical reporting found across fields may reduce our ability to confidently build on published works across disciplines, as well as within them. Since researchers from different disciplines use the same methods, it would be reasonable to agree on common standards for statistical testing and reporting. Developing cross-disciplinary standards on statistical testing and reporting will improve the scientific progress and increase the confidence in research results.

To better understand the differences between disciplines, we approached 11 researchers from economics (5), computer science (5), and management (1) with a question when will they test and report assumptions for the methods we explored in our study. To demonstrate how different are the standards lets focus on the previously discussed normality assumption for t-test. According to our survey, only few economists test the normality assumption and none of them expect this assumption to be reported. As one of them said "*I always plot the distribution of my variables before running formal tests such as unpaired t-tests. If I run an unpaired t-test I would plot 2 separate distributions. However, I would rarely run formal tests for normality*." Another economist researcher stressed that this assumption is rarely addressed due to a big sample of typical economics studies. On the other hand, computer scientists reported that they test the normality assumption, but only roughly half of them report about this and do not expect other researchers to report on this. This simple example confirms that lack of reporting might be not due to lack of testing but because it is assumed to be self-evident that authors tested assumptions. Yet, we argue that this lack of transparency is not beneficial neither for reviewers nor for the broad audience of readers.

In addition, our results demonstrate an overall very low prevalence of reporting of statistical assumptions. While this is not necessarily the first, or the only problem that needs to be solved to improve the validity and rigor of scientific work, we argue that it is an important component of the solution. A better practice when it comes to statistical assumption reporting would require scientists to make their assumptions explicit. Paired with a requirement to publish one's data this would hopefully lead to more reproducible and reliable statistical analyses in scientific papers.

10. References

Affairs, L.W. and the T.F. on S.I. a P. a B. of S., 1999. Statistical methods in psychology journals. *American psychologist*, 54 (8)(8), pp.594–604.

Agrawal, A. et al., 2013. Digitization and the Contract Labor Market: A Research Agenda. *NBER Working Paper*, p.37.

Alasuutari, P., 2010. The rise and relevance of qualitative research. *International Journal of Social Research Methodology*, 13(2), pp.139–155.

Anderson, C., 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 16(7), pp.1–2. Available at: http://www.wired.com/print/science/discoveries/magazine/16-07/pb_theory.

Baker, M., 2016a. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), pp.452–454. Available at: http://www.nature.com/doifinder/10.1038/533452a.

Baker, M., 2016b. Statisticians issue warning on Pvalues. *Nature*, 531, p.151.

BARNES, W.H.F., 1944. The Nature of Explanation. *Nature*, 153(3890), pp.605–605. Available at: http://www.nature.com/articles/153605a0.

Bernstein, A., 2000. How can cooperative work tools support dynamic group process? Bridging the specificity frontier. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. pp. 279–288.

Bernstein, A., Klein, M. & Malone, T.W., 2012. Programming the global brain. *Communications of the ACM*, 55(5), p.41.

Bernstein, M.S. et al., 2010. Soylent: a word processor with a crowd inside. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp.313–322.

de Boer, P.M. & Bernstein, A., 2015. PPLib: Towards the Automated Generation of Crowd Computing Programs using Process Recombination and Auto-Experimentation. *ACM Transactions on Intelligent Systems and Technology*, (Special Issue: Crowd Computing).

Bollier, D., 2010. *The Promise and Peril of Big Data*, Available at: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.

Boyd, D. & Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, 15(5), pp.662–679.

Campbell, A. & Wu, A.S., 2011. Multi-agent role allocation: Issues, approaches, and multiple perspectives. *Autonomous Agents and Multi-Agent Systems*, 22(2), pp.317–355.

Campbell, J.L. et al., 2013. Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods and Research*, 42(3), pp.294–320.

Carpenter, J., 2011. May the best analyst win. *Science (New York, N.Y.)*, 331(6018), pp.698–699.

Chi, M.T.H., 2008. Three Types of Conceptual Change: Belief Revision, Mental Model Transformation, and Categorical Shift. In *Handbook of research on conceptual change*. pp. 61–82.

Collaboration, O.S., 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251), p.aac4716-aac4716. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716.

Conklin, E.J. & Yakemovic, K.C.B., 1991. A Process-Oriented Approach to Design Rationale. *Human-Computer Interaction*, 6, pp.357–391.

Creswell, J., 2002. *Educational research: Planning, conducting, and evaluating quantitative*

and qualitative research,

Davenport, T.H. & Patil, D.J., 2012. Data_Scientist-the_Sexiest_Job_of_the_21St_Century.Pdf. , pp.70–76.

Davey Smith, G. & Ebrahim, S., 2002. Data dredging, bias, or confounding. *Bmj*, 325(7378), pp.1437–1438. Available at: http://www.bmj.com/cgi/doi/10.1136/bmj.325.7378.1437.

Van Dijck, J. & Nieborg, D., 2009. Wikinomics and its discontents: a critical analysis of Web 2.0 business manifestos. *New Media & Society*, 11(5), pp.855–874. Available at: http://journals.sagepub.com/doi/10.1177/1461444809105356.

Dissanayake, I., Zhang, J. & Gu, B., 2014. Virtual Team Performance in Crowdsourcing Contests : A Social Network Perspective. *ICIS 2015 Proceedings*, (Savage 2012), pp.1–16.

Dwork, C. et al., 2015. validity in adaptive data analysis. *Science*, 349(6248), pp.636–638. Available at: http://www.sciencemag.org/content/349/6248/636.

Edgell, S.E. & Noon, S.M., 1984. Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin*, 95(3), pp.576–583.

Eicken, H., 2013. Six red flags for suspect work. *Nature*, 497, pp.433--434.

Erceg-Hurn, D.M. & Mirosevich, V.M., 2008. Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), pp.591–601. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.63.7.591.

Fahy, P., 2001. Addressing some common problems in transcript analysis. *The International Review of Research in Open and Distance Learning*, 1(2), pp.1–6. Available at: http://www.irrodl.org/index.php/irrodl/article/viewArticle/321.

Feldman, M., Juldaschewa, F. & Bernstein, A., 2017. Data Analytics on Online Labor Markets: Opportunities and Challenges. Available at: http://arxiv.org/abs/1707.01790 [Accessed August 12, 2017].

Feldman, Mi., Anastasiu, C. & Bernstein, M., 2016. Towards Enabling Crowdsourced Collaborative Data Analysis. *Collective Intelligence*, (June), pp.1–5.

Fernandes-Taylor, S. et al., 2011. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC research notes*, 4(1), p.304. Available at: http://www.biomedcentral.com/1756-0500/4/304 [Accessed July 29, 2015].

Field, A., 2013. Discovering Statistics using IBM SPSS Statistics. *Discovering Statistics using IBM SPSS Statistics*, pp.297–321.

Fiske, S.T., 2016. How to publish rigorous experiments in the 21st century. *Journal of Experimental Social Psychology*, 66, pp.4–6. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022103116000032.

Fox, P. & Hendler, J., 2011. Changing the equation on scientific data visualization. *Science*, 331(6018), pp.705–708.

Friedkin, N.E. et al., 2016. Network science on belief system dynamics under logic constraints. *Science*, 354(6310), pp.321–326.

Gelman, A. & Hennig, C., 2015. Beyond subjective and objective in statistics. *arXiv preprint arXiv:1508.05453*. Available at: http://arxiv.org/abs/1508.05453 [Accessed September 16, 2016].

Gelman, A. & Loken, E., 2014a. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Psychological bulletin*, 140(5), pp.1272–1280. Available at: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf%5Cnhttp://doi.apa.org/getdoi.cfm?doi=10.1037/a0037714.

Gelman, A. & Loken, E., 2014b. The statistical Crisis in science. *American Scientist*, 102(6), pp.460–465.

Gelman, A. & Shalizi, C.R., 2015. Philosophy and the practice of Bayesian statistics Andrew. *British Journal of Mathematical and Statistical Psychology*, 66(1), pp.8–38.

Gilad-Bachrach, R., Navot, A. & Tishby, N., 2004. Margin based feature selection - theory and algorithms. In *Proceedings of the 21st International Confer- ence on Machine Learning*. p. 43. Available at: http://eprints.pascal-network.org/archive/00000869/.

Glaser, B.G. & Strauss, A.L., 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Available at: http://www.amazon.com/dp/0202302601.

Glass, G. V, Peckham, P.D. & Sanders, J.R., 2012. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance Author (s): Gene V . Glass , Percy D . Peckham and James R . Sanders Reviewed work (s): Source : Review of Educational Research , Vol . 42 , N. *Review of Educational Research*, 42(3), pp.237–288.

Good, P.I. & Hardin, J.W., 2012. *Common errors in statistics (and how to avoid them)*, John Wiley & Sons.

Gregor, S., 2006. The nature of theory in information systems. *MIS Quartely*, 30(3), pp.611–642.

Grolemund, G. & Wickham, H., 2014. A Cognitive Interpretation of Data Analysis. *International Journal of Statistics*, 82(2), pp.184–204. Available at: http://vita.had.co.nz/papers/sensemaking.pdf%5Cnhttp://onlinelibrary.wiley.com/doi/10.1111/insr.12028/abstract.

Gruber, T.R. & Russell, D.M., 1993. Generative Design Rationale: Beyond the Record and Replay Paradigm. *Design rationale: Concepts*, (December 1993). Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.1981&rep=rep1&type=pdf%5Cnpapers2://publication/uuid/AEA45CFD-89DF-4DE4-BC2D-71283E2E5DFB.

Guindon, R., 1990. Knowledge exploited by experts during software system design. *International Journal of Man-Machine Studies*, 33(3), pp.279–304.

Gutierrez, D.D., 2015. *Machine learning and data science: an introduction to statistical learning methods with R*, echnics Publications.

Haas, D. et al., 2015. Wisteria : Nurturing Scalable Data Cleaning Infrastructure. *Proceedings of the 41st International Conference on Very Large Data Bases*, 8(12), pp.2004–2007.

Head, M.L. et al., 2015. The Extent and Consequences of P-Hacking in Science. *PLoS*

Biology, 13(3).

Heer, J., Viégas, F.B. & Wattenberg, M., 2009. Voyagers and Voyeurs: Supporting Asynchronous Collaborative Visualization. *Communications of the ACM*, 52(1), pp.87–97.

Hevner, A.R. et al., 2004. Design Science in Information Systems Research. *MIS quarterly*, 28(1), pp.75–105.

Hill, R.C. & Levenhagen, M., 1995. Metaphors and Mental Models: Sensemaking and Sensegiving in Innovative and Entrepreneurial Activities. *Journal of Management*, 21(6), pp.1057–1074.

Hoekstra, R., Kiers, H.A.L. & Johnson, A., 2012. Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3(MAY), pp.1–9.

Howison, J. & Crowston, K., 2013. Olla boration through open superposition.

Hruschka, D.J. et al., 2004. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods*, 16(3), pp.307–331. Available at: http://journals.sagepub.com/doi/10.1177/1525822X04266540.

Humphreys, M., Sanchez de la sierra, R. & Van der windt, P., 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1), pp.1–20.

Intelligence, A., 1992. Task-Structure Analysis Knowledge MOdeling for. , 35(9).

Introne, J. et al., 2013. Solving wicked social problems with socio-computational systems. *Kuntsliche Intelligenz*, 27(1), pp.45–52. Available at: http://cci.mit.edu/working_papers_2012_2013/cciwp2012-05colabkunstinel.pdf.

Ioannidis, J.P.A., 2005. Why most published research findings are false. *PLoS Medicine*, 2(8), pp.0696–0701.

Johnson-Laird, P.N., 1980. Mental models in cognitive science. *Cognitive Science*, 4(1), pp.71–115.

Jussim, L. et al., 2015. Interpretations and methods: Towards a more effectively self-correcting social psychology ☆. *Journal of Experimental Social Psychology*, xxx, pp.116–133. Available at: http://dx.doi.org/10.1016/j.jesp.2015.10.003.

Kalleberg, A.L. & Dunn, M., 2016. Good Jobs, Bad Jobs in the Gig Economy. *The Gig Economy: Employment Implications: Perspectives on Work 2016*, 20, pp.10–14.

Kandel, S. et al., 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. *Human factors in computing systems. ACM*, pp.3363–3372.

Kanji, G. k, 2006. *100 Statistical Tests* 3rd ed., London: SAGE Publications India Pvt Ltd. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006199-199501000-00015.

Kaptein, M. & Robertson, J., 2012. Rethinking Statistical Analysis Methods for CHI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.1105–1114. Available at: http://doi.acm.org/10.1145/2207676.2208557.

Kay, M., Nelson, G.L. & Hekler, E.B., 2016. Researcher-Centered Design of Statistics.

Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16, (August), pp.4521–4532. Available at: http://dl.acm.org/citation.cfm?doid=2858036.2858465.

Kittur, A. et al., 2011. CrowdForge: Crowdsourcing Complex Work. *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, pp.43–52. Available at: http://dl.acm.org/citation.cfm?doid=2047196.2047202.

Kittur, A. et al., 2012. CrowdWeaver : Visually Managing Complex Crowd Work. *Scenario*, pp.1033–1036. Available at: http://www.cs.cmu.edu/~pandre/pubs/crowdweaver-cscw2012.pdf.

Kittur, A., Nickerson, J. & Bernstein, M., 2013. The Future of Crowd Work. *Proc. CSCW '13*, pp.1–17. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2190946%5Cnpapers2://publication/uuid/AE6BF263-1DEF-4900-8C95-DC8BAD2DE4AF.

Klein, G. & Moon, B., 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5), pp.88–92.

Klein, R.A. et al., 2014. Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), pp.142–152.

Krippendorff, K., 2004. *Content Analysis: An Introduction to Its Methodology*,

Krishnan, S. et al., 2015. SampleClean: Fast and Reliable Analytics on Dirty Data. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pp.59–75. Available at: http://sites.computer.org/debull/A15sept/p59.pdf.

Kulkarni, A., Can, M. & Hartmann, B., 2012. Collaboratively crowdsourcing workflows with turkomatic. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, p.1003. Available at: http://dl.acm.org/citation.cfm?doid=2145204.2145354.

Kurasaki, K.S., 2000. Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data. *Field Methods*, 12(3), pp.179–194.

Kuzon, W., Urbanchek, M.G. & Mccabe, S.J., 1997. Seven deadly sins of statistical analysis. *Journal of Oral and Maxillofacial Surgery*, 55(8), pp.897–898. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0278239197903773.

Lang, T. a & Altman, D.G., 2013. Basic Statistical Reporting for Articles Published in Biomedical Journals : The " Statistical Analyses and Methods in the Published Literature " or The SAMPL Guidelines ". *Science editors' handbook*, pp.29–32. Available at: http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf.

Langlois, R.N., 2002. Modularity in technology and organization. *Journal of Economic Behavior and Organization*, 49(1), pp.19–37.

Lee, J. & Lai, K.Y., 1991. What's in Design Rationale? *Humanï¿½Computer Interaction*, 6(3–4), pp.251–280.

Leek, J.T. & Peng, R.D., 2015. P values are just the tip of the iceberg. *Nature*, 520(7549), p.612.

Lukacs, P.M., Burnham, K.P. & Anderson, D.R., 2010. Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, 62(1), pp.117–125.

MacDonald, J., 2003. Assessing online collaborative learning: Process and product. *Computers and Education*, 40(4), pp.377–391.

MacLean, A. et al., 1991. Questions, Options, and Criteria: Elements of Design Space Analysis. *Humanï¿½Computer Interaction*, 6(3–4), pp.201–250.

Malone, T.W. et al., 1999. Tools for Inventing Organizations : Toward a Handbook of Organizational Processes Tools for Inventing Organizations : Toward a Handbook of Organizational Processes. , 3(May 2015), pp.425–443.

Mann, M., 2016. Must try harder. *New Scientist*. Available at: http://www.sciencedirect.com/science/article/pii/S0262407916303682 [Accessed August 30, 2016].

Martin Bland, J. & Altman, D., 1986. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet*, 327(8476), pp.307–310.

Mascha, E.J., 2010. Equivalence and noninferiority testing in anesthesiology research. *Anesthesiology*, 113(4), pp.779–781.

Miles, M., Huberman, M. & Saldana, J., 2014. *Qualitative Data Analysis*,

Morton, K. et al., 2014. Support the Data Enthusiast: Challenges for Next-Generation Data-Analysis Systems. *Proceedings of the VLDB Endowment, Volume 7, pp. 453–456, 2014*, 7, pp.453–456. Available at: http://homes.cs.washington.edu/~kmorton/p446-morton.pdf.

Nimon, K.F., 2012. Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, 3(AUG), pp.1–5.

Van Noorden, R., 2014. Online collaboration: Scientists and the social network. *Nature*, 512(7513), pp.126–129. Available at: http://www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711.

Norman, D.A., 1983. Some Observations on Mental Models. In *Mental Models*. pp. 7–14. Available at: http://www.amazon.com/Mental-Models-Cognitive-Science-Series/dp/0898592429.

Nosek, B.A., Spies, J.R. & Motyl, M., 2012. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, 7(6), pp.615–631. Available at: http://pps.sagepub.com/lookup/doi/10.1177/1745691612459058.

Nuzzo, R., 2014. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487), pp.150–152.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251), p.aac4716-aac4716. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26315443.

Osborne, J. & Waters, E., 2002. Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, 8(2), p.1.

Ott, E.M., 1989. Effects of the Male-Female Ratio at Work: Policewomen and Male Nurses. *Psychology of Women Quarterly*, 13(1), pp.41–57.

Paglieri, F., 2004. Data-oriented belief revision: Towards a unified theory of epistemic processing. *Proceedings of STAIRS*. Available at: http://books.google.com/books?hl=en&lr=&id=Z569jqwQuK8C&oi=fnd&pg=PA179&dq=Data-oriented+Belief+Revision+:+Towards+a+Unified+Theory+of+Epistemic+Processing&ots=SqAEHHjdec&sig=Out0eaWHx3vygt_eBCzHm6QJYcA.

Partington, D., 2013. *Essential Skills for Management Research*,

Peffers, K. et al., 2008. A Design Science Research Methodology for Information Systems Research. *J. Manage. Inf. Syst.*, 24(January), pp.45–77.

Ransbotham, S., Kiron, D. & Prentice, P.K., 2015. The Talent Dividend. *MIT Sloan Management Review*, 56(4), pp.1–12. Available at: http://sloanreview.mit.edu/projects/analytics-talent-dividend/.

Redmiles, D., 2000. Software Requirements for Supporting Collaboration through Categories.

Reinecke, K. & Bernstein, A., 2013. Knowing What a User Likes: A Sesign Science Approach to Interfaces that Automatically Adapt to Culture. , 37(2), pp.427–453.

Rouder, J.N. et al., 2016. Is There A Free Lunch In Inference? *topiCS*, 8(1), pp.1–5.

Russell, D.M. et al., 1993. The cost structure of sensemaking. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93*. pp. 269–276. Available at: http://portal.acm.org/citation.cfm?doid=169059.169209.

Russo, D. & Zou, J., 2016. Controlling Bias in Adaptive Data Analysis Using Information Theory. In *Proceedings of the 19th International Con- ference on Artificial Intelligence and Statistics (AISTATS) 2016*. pp. 1232–1240. Available at: http://arxiv.org/abs/1511.05219.

Saldana, J., 2011. *Fundamentals of Qualitative Research: Understanding Qualitative Research*,

Salehi, N. et al., 2016. Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

dos Santos, F. & Bazzan, A.L.C., 2009. An ant based algorithm for task allocation in large-scale and dynamic multiagent scenarios. *Proceedings of the 11th Annual conference on Genetic and evolutionary computation - GECCO '09*, p.73. Available at: http://portal.acm.org/citation.cfm?doid=1569901.1569912.

Van Schaik, P. & Weston, M., 2016. Magnitude-based inference and its application in user research. *International Journal of Human Computer Studies*, 88(August), pp.38–50.

Schlauderer, S. & Overhage, S., 2013. Exploring the Customer Perspective of Agile Development: Acceptance Factors and on-Site Customer Perceptions in Scrum Projects. *Thirty Fourth International Conference on Information Systems*, pp.1–20.

Schubanz, M., 2014. Design rationale capture in software architecture: What has to be captured? In *WCOP 2014 - Proceedings of the 19th International Doctoral Symposium on Components and Architecture (Part of CompArch 2014)*. pp. 31–36. Available at: http://dx.doi.org/10.1145/2601328.2601329.

Sculley, D. & Pasanek, B.M., 2008. Meaning and mining: The impact of implicit

assumptions in data mining for the humanities. *Literary and Linguistic Computing*, 23(4), pp.409–424.

Seel, N.M., 2001. Epistemology, situated cognition, and mental models: "Like a bridge over troubled water." *Instructional Science*, 29(4–5), pp.403–427.

Seitz, F., Heisenberg, W. & Pauli, W., 2000. Decline of the generalist The vigour of every discipline depends on people of broad vision . *Nature*, 403(February), pp.10021–10021.

Sere, F.C. et al., 2011. Online collaboration: Collaborative behavior patterns and factors affecting globally distributed team performance. *Computers in Human Behavior*, 27(1), pp.490–503.

Sheskin, D.J., 2004. Handbook of parametric and nonparametric statistical procedures. *Technometrics*, 46, p.1193. Available at: http://books.google.com/books?id=bmwhcJqq01cC&pgis=1.

Silberzahn, R. & Uhlmann, E.L., 2015. Many Hands Make Tight Work. *Nature*, 526(7572), pp.189–191. Available at: http://www.nature.com/news/crowdsourced-research-many-hands-make-tight-work-1.18508.

Simmons, J.P., Nelson, L.D. & Simonsohn, U., 2011. False-Positive Psychology. *Psychological Science*, 22(11), pp.1359–1366. Available at: http://journals.sagepub.com/doi/10.1177/0956797611417632.

Smith, A.J., 1990. The Task of the Refereee. , pp.1–7.

Stefik, M., 1981. Planning with constraints (MOLGEN: Part 1). *Artificial Intelligence*, 16(2), pp.111–139.

Stein, R.T. & Heller, T., 1979. An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *Journal of Personality and Social Psychology*, 37(11), pp.1993–2002.

Strasak, A.M. et al., 2007. Statistical errors in medical research - A review of common pitfalls. *Swiss Medical Weekly*, 137(3–4), pp.44–49.

Strauss, A. & Corbin, J., 1990. Basics of qualitative research: grounded theory procedure and techniques. *Qualitative Sociology*, 13(1), pp.3–21.

Thomas, D.R., 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2), pp.237–246. Available at: http://journals.sagepub.com/doi/10.1177/1098214005283748.

Tseng, H. et al., 2009. Key Factors in Online Collaboration and Their Relationship to Teamwork Satisfaction. *The Quarterly Review of Distance Education*, 10(626), pp.195–206.

Tukey, J.W. & Wilk, M.B., 1966. Data analysis and statistics: an expository overview. *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference*, (695), pp.695–710. Available at: http://dl.acm.org/citation.cfm?id=1464366.

Vargha, A. & Delaney, H., 1998. The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, 23(2), pp.170–192.

Viegas, F.B. et al., 2007. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), pp.1121–1128.

Weiss, G. & Wodak, R., 2003. *Critical Discourse Analysis*,

Westfall, J. & Yarkoni, T., 2016. Statistically controlling for confounding constructs is harder than you think. *PloS one*, 11(3), pp.1–22.

Willett, W. et al., 2011. CommentSpace: Structured Support for Collaborative Visual Analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.3131–3140. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.188.1845&rep=rep1&type=pdf.

de Winter, J.C.F. & Dodou, D., 2010. Five-Point Likert Items : t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11), pp.1–16. Available at: http://pareonline.net/pdf/v15n11.pdf.

Woolston, C., 2015. Psychology journal bans P values. *Nature*, 519(7541), pp.9–9. Available at: http://www.nature.com/doifinder/10.1038/519009f [Accessed August 12, 2017].

Yadav, M.S. & Pavlou, P.A., 2014. Marketing in Computer-Mediated Environments: Research Synthesis and New Directions. *Journal of Marketing*, 78(1), pp.20–40. Available at: http://journals.ama.org/doi/abs/10.1509/jm.12.0020.

Yukl, G., 2001. Leadership in organizations. *Personnel Psychology*, 7th(4), p.542. Available at: http://files.liderancaecoaching.webnode.com/200000015-31f5732fb3/media-F7B-97-randd-leaders-business-yukl.pdf.

Zimmerman, D.W., 2004. Inflation of Type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation tests. *Psicologica*, 25(1), pp.103–133.

Zimmerman, D.W., 1998. Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. *The Journal of Experimental Education*, 67(1), pp.55–68.

11. Appendices

11.1 Methods

A list of methods ordered by their popularity

	Methods
1	MANOVA
2	ANOVA
3	ANCOVA
4	Linear regression
5	Logistic regression
6	Correlation
7	U test
8	Wilcoxon Signed-Ranks test
9	Kruskal-Wallis
10	McNemar
11	Friedman's test
12	chi square
13	t-test
14	Dunn's test
15	Receiver Operating Characteristic Curve
16	Odds ratio
17	Kappa
18	Survival analysis
19	Sensitivity
20	Post hoc analysis
21	Cluster analysis
22	Factor analysis
23	Classification and regression tree analysis
24	Mantel Haenszel
25	Decision tree analysis
26	Meta-analysis
27	Content analysis
28	Confidence interval
29	Random effect
30	Cronbach's alpha

Data Analytics on Online Labor Markets: Opportunities and Challenges

This chapter is based on a paper that is currently under review at the Data Science Journal**.** The authors contributing to this chapter are: Michael Feldman, Frida Juldaschewa, Abraham Bernstein.

Data Analytics on Online Labor Markets: Opportunities and Challenges

Abstract

The data-driven economy has led to a significant shortage of data scientists. To address this shortage, this study explores the prospects of outsourcing data analysis tasks to freelancers available on online labor markets (OLMs) by identifying the essential factors for this endeavor. Specifically, we explore the skills required from freelancers, collect information about the skills present on major OLMs, and identify the main hurdles for out-/crowd-sourcing data analysis. Adopting a sequential mixed-method approach, we interviewed 20 data scientists and subsequently surveyed 80 respondents from OLMs. Besides confirming the need for expected skills such as technical/mathematical capabilities, it also identifies less known ones such as domain understanding, an eye for aesthetic data visualization, good communication skills, and a natural understanding of the possibilities/limitations of data analysis in general. Finally, it elucidates obstacles for crowdsourcing like the communication overhead, knowledge gaps, quality assurance, and data confidentiality, which need to be mitigated.

1. Introduction

In the past years it has become evident, that there is a continuously growing demand for data scientists and for people who are able to systematically interpret data. Since the availability of data is growing faster than the availability of experts with the relevant skillset to interpret it, finding competent experts for data analysis tasks is becoming increasingly challenging due to a variety of required skills. Contrary to the past, when benefiting from data products was a prerogative of big companies, having server farms and in-house teams of support technicians, currently, the development of cloud computing allows for on-demand usage of computational resources at reasonable costs. Therefore, even small and medium sized enterprises (SMEs) have started collecting data about their customers, business transactions, and other records related to their business. However, analysis of the collected data is hampered by the growing shortage of data experts capable of analyzing the data and producing comprehensive insights that are intelligible to a wide audience of population and decision makers (Davenport and Patil 2012).

Often, lacking internally available talent, companies are compelled to seek for external solutions that will allow to make sense of their data. Recognizing the need for on-demand data analysis, multiple software companies, such as Microsoft and Google, started developing and offering cloud-based data analytics products that allow running various machine learning (ML) algorithms on a big scale in the cloud. Some of these products claim to reduce the complexity threshold of data analysis by allowing to plug the data to their services and subsequently to run data analysis as a black box process. However, this approach leads to questionable results given the limited control over the data analysis process and reduced flexibility to preprocess the data and to tailor the models for specific needs. In addition, statistical expert systems (Serban et al. 2013) have been proposed to address the matter. Many of these systems require data analysts in order to be employed correctly. Therefore, even though data analysis has been made much more accessible by this variety of

tools, there is an apparent need for skilled experts to conduct and orchestrate the process of data analysis.

The lack of experts available in geographical proximity can be resolved by online labor markets (OLM), that overcome multiple drawbacks and allow to hire experts in a flexible manner. The potential of such platforms has been recognized by the industry and the earnings of such platforms experienced sharp increase during past years, attracting growing numbers of freelancers and employers. Unfortunately, research on the necessary skills for potentially crowdsourced data analysis still remains in a blind spot and has not been sufficiently addressed yet. Therefore, we carry out a study to both analyze the requirements and explore the skills available online. We argue for the importance of this phenomenon and hope that our work will promote the discussion on the major constructs composing the candidate selection process on OLM.

Various data analysis tasks require diverse knowledge and skills. While some tasks are fairly straightforward and effortless, others require proficiency in multiple disciplines and hands-on experience. To illustrate, descriptive statistics require a somewhat shallow statistical background while other methods, such as neural networks or support vector machine learning frequently require in-depth understanding. Therefore, our first research question is: ***RQ1*** *What are the skills required in data analysis?* Identifying the skills required for a certain task is the first step in finding suitable workers to perform it. For the successful assignment of freelancers to tasks we also need to understand the capabilities of the workers. Hence, gaining an insight about the skills of freelancers will allow to design tasks while taking into account the constraints imposed by the existing talent pool in the online labor market. Therefore, the second research question is: ***RQ2*** *What are the relevant skills and characteristics that freelancers in OLM possess, and do they match the required skills for data analysis (identified for RQ1)?* The last research question deals with various obstacles to the outsourcing of data analysis to freelancers and gain insights on the potential ways to resolve these problems. Hence, our third research question is: ***RQ3*** *What are the obstacles to outsourcing data analysis to OLM?*

To answer these questions, we conduct an exploratory study including 20 interviews with data scientists, followed by a survey with 80 freelancers to learn about the talent available on major freelance platforms. Our contribution is twofold: 1) we fill the research gap concerning the endeavor of outsourcing data analysis to OLM by means of systematic research on needs vs. supplies of explicit skills and 2) provide the necessary basis for future theorization on employee selection in an online setting.

The remainder of this paper is structured as follows: In the next section we explore the related work in the domains of crowdsourcing as well as online labor markets and give an overview of the theoretical work related to our study. Then we describe our research approach and outline the qualitative and quantitative results of our study. Next, we discuss the results of our study and present limitations and future research. Finally, we conclude with a summary of the paper.

2. Related Work

Crowdsourcing has gained increased relevance and acceptance as an approach for outsourcing activities to an online community. In recent years, the business

community embraced the approach of outsourcing some activities to a crowd by means of evolved and specialized web-based platforms. Jobs are mostly partitioned into groups of simplified sub-tasks and distributed to crowd workers in an open call manner (Howe 2008; Alam and Campbell 2014). Even though crowdsourcing has a long history (just consider the Longitude prize or the Oxford English Dictionary), its current popularity is largely accounted to Amazon's establishment of the first crowdsourcing Internet marketplace called Mechanical Turk (MTurk).[18] This platform provides a wealth of paid micro-tasks that require minimal time and cognitive effort, but aggregates results in major accomplishments (Kamar et al. 2012). As for today, MTurk has preserved its leading role as the most popular platform for micro-tasks. However, the phenomenon of online work has expanded to include platforms supporting laborious expert work in various domains. Platforms such as Upwork and Freelancer,[19] offer the service of matchmaking between employers and freelancers based on reported expertise. Other platforms such as TopCoder, 99designs, or Kaggle[20] offer contest-based participation, while yet other platforms like InnoCentive or Idea Bounty[21] are crowdsourcing ideas for solving challenging problems. Encouraged by the potential of human computation, recently, attempts have been made to extend human computation tasks beyond relatively simple and non-demanding ones (e.g. Haas et al. 2015). As a consequence, the crowdsourcing domain is faced with an emerging need for concepts and paradigms for the assignment of complex tasks that require a wide spectrum of human abilities and talents to suitable crowd workers. Finding appropriate crowd workers, is non-trivial due to the motivational, cognitive, and error diversity of humans (Bernstein et al. 2012). Moreover, the remote and unstable character of most crowd markets limits the ability to track and profile workers and gives rise to an even greater challenge of establishing trustworthy and robust recruiting policies (Ipeirotis 2010).

Evidently, platforms such as MTurk are primarily designed for micro-tasks and do not support complex and ill-defined tasks that require well-established coordination and trust relationships between the requesters and crowd workers. The rise of freelancer platforms aims to fill this gap by supporting both, employers and crowd workers/freelancers, with a user interface and workflow that enables to conduct complex projects in a remote manner. In contrast to the micro-task labor market that has reached saturation and even some decline (see Ipeirotis' analysis on MTurk[22]), the online freelancer market has grown substantially during the last years. It is now bringing together millions of freelancers and employers (Agrawal et al. 2013). This shift exemplifies the transition of online labor markets from simple, short-term, and low effort jobs, as they were originally common in online labor markets, to complex and long-term tasks that are typical to traditional work settings. A growing number

[18] www.mturk.com/mturk/welcome

[19] www.upwork.com, www.freelancer.com

[20] www.topcoder.com, www.99designs.com, www.kaggle.com

[21] www.innocentive.com, www.ideabounty.com

[22] http://www.behind-the-enemy-lines.com/2016/02/a-cohort-analysis-of-mechanical-turk.html

of white collar workers are switching to online labor markets due to the advantages online work can offer. Students, homemakers, retired experts, and single parents are frequently found on OLM and the prevailing professions are graphic design, copywriting, data entry, and programming (Mill 2011).

Similarly to the general shortage of data scientists, i.e. experts able to provide comprehensive data-driven solutions, (Davenport and Patil 2012), it is fairly uncommon to find data analysis experts on a crowdsourcing platform. It is, however, not unusual to find workers possessing some partial knowledge and/or willing to learn a new topic. A survey of 153 data scientists, conducted by Crowdflower (2015), has revealed that data scientists mostly refer to themselves as researchers (54%), computer scientists (52%), BI analysts (36%), and mathematicians (19%). Additionally, most of the respondents mentioned they are working with Excel (56%), R (43%), and Tableau (26%). The majority of the respondents consider data cleaning and organizing as the most time consuming task (67%), while 53% say that collecting data sets is the most laborious. This matches with Kurgan and Musilek (2006) who surveyed multiple papers evaluating the relative effort in different data analysis activities and concluded that data preparation is by far the most time consuming activity with estimates ranging between 45 and 60%. The most demanded skills are programming and statistics and proficiency in Python and R are by far the most prominent. According to Harris et al. (2013), a data scientist has to be capable of 1) designing statistical models, 2) creating machine learning and text mining algorithms, 3) cleaning and converting raw data, 4) carrying out quality assurance testing to ensure the quality of models, and 5) communicating the results through clear data visualization. Other supplementary skills such as communication, collaboration, and creativity are also mentioned as key success factors. Chatfield et al. (2014) have analyzed a body of literature from six major academic databases and derived a set of six data science attributes: 1) Entrepreneurship and business domain knowledge, 2) Computer scientist, 3) Effective Communication skills, 4) Create valuable and actionable insights, 5) Inquisitive and curious, and 6) Statistics and modeling.

Evaluating skills of job candidates is one of the major challenges in both online and traditional labor markets. Even though some platforms allow job candidates to perform various online tests to assess their competence in a variety of topics, cheating and tests' leakage hamper reliable evaluation. Moreover, technological advancement requires tests to be frequently updated and reliably evaluated, promising that the performance on these tests adequately reflects a candidate's skills (Christoforaki and Ipeirotis 2015). One major difference of online compared to traditional labor markets is the highly heterogeneous workforce composed of a crowd with different skills spanning across a variety of different professions. While the candidates in traditional markets share some similarities such as common cultural and geographical background, candidates in online labor markets come from all around the world and exhibit high variance in qualities and skills. Kokkodis and Ipeirotis (2015) assume these skills to be latent, however, possible to be measured through the worker's available characteristics on a platform such as employee ratings, accomplished projects, hours worked, and wages. Utilizing these characteristics, they present a number of models that estimate the workers' latent skills and their evolution over time. Feldman and Bernstein (2014) propose that

cognitive abilities of freelancers are another latent factor that, to a large extent, predefines the performance on various crowd-tasks. They examine the performance on various crowd-tasks with different setups to predict task performance where cognitive abilities, performance on previous crowd-tasks, or both are partially known, and show that cognition-based task assignment leads to an improvement in task performance prediction. Suzuki et al. (2016) propose to support the skill development of workers by introducing a concept of micro-internships. According to the proposed solution, micro-internships allow workers to learn, improve, and develop new skills. At the same time employers can evaluate the skills of a candidate. Verroios et al. (2015) are grouping employers on Odesk (today Upwork) with respect to their hiring criteria and learn the hiring preferences for each cluster. Results show that while some groups of clients are positively biased to freelancers that are new to a marketplace, others ignore their reputation and focus primarily on a person's job fit. In this context, Pallais (2013) uses a field experiment in Odesk to show that awarding new freelancers with a first job benefits the market with information about their abilities and increases freelancers' average earnings.

The aforementioned studies provide a high level understanding of who online freelancers are and what data analysis tasks that can be outsourced. They, however, do not shed light on what concrete, well-specified skills are required and whether they are available on OLM.

3. Research Approach

Our study adopts a sequential mixed-method approach, harnessing both the power of qualitative as well as of quantitative research (Teddlie and Tashakkori 2009). The qualitative study comprises an interpretive case study, in which 20 data analysis experts are interviewed (Walsham 2006). The quantitative study consists of a web survey (Dillman 2011) following a descriptive and cross-sectional research design as it has been outlined by Pinsonneault and Kraemer (1993). The results of the first research phase inform the design of the second research phase. Using such a multiple-method approach leads to higher robustness of results due to triangulation – leveraging the usage of multiple methods, data sources, or theories to facilitate deeper understanding of the phenomenon (Denzin 1973). Our approach allows on the one hand for data triangulation, the use of a variety of sources in a study, and on the other hand for methodological triangulation, the use of multiple methods to study a research problem. As qualitative research is especially appropriate for studying complex phenomena, we applied this method first to explore the domain of data scientists and to gain in-depth descriptions and understanding of their environment. After having gained a thorough understanding, quantitative research is well suited to apply the learned content onto a broader population and obtain quantitative data to generalize research findings (Johnson and Onwuegbuzie 2004). Overall, our process consists of four steps: first, qualitative data is collected through interviews with data scientists and data analysts. Second, the qualitative data is used to identify necessary skills and other factors for outsourcing data analysis tasks to freelancers. Third, an online survey is designed based on the previously identified skills and distributed on various freelancing platforms. Lastly, the generated quantitative data is analyzed to compare the skills available on online labor markets with the desired skills.

3.1 Qualitative research

We approached potential participants for interviews through personal contacts, professional online business networks (e.g. LinkedIn, XING, Data Science Central),[23] and professional meetups[24]. By exploring various sources, we intended to compose a sample of individuals with diverse backgrounds, spanning different industries and positions. The main prerequisite for participation in our study was the profession criterion; specifically, we aimed at individuals who either hold positions of data scientists or data analysts, or are primarily occupied with data analysis in their daily work. Whereas job titles and job requirements vary throughout organizations, the fact that 'a data scientist represents an evolution from the business or data analyst role' (Zikopoulos et al. 2012) suggests that they have a common foundation and work with data to answer business questions (Kandel et al. 2012). Almost all interviewees graduated from quantitative disciplines with educational degrees as follows: 1 Post-doctorand, 4 Ph.D., 12 M.Sc. and 3 with Bachelor degree. They are employed in digital analytics, insurance, financial services, analytics consultancies, retail, telecommunication and internet broadcasting with median of 6 years experience. Therefore, all selected participants are experts in data analysis, and thus able to provide valuable insights about the domain of data analysis. In total, 20 semi-structured interviews, lasting between 30 and 60 minutes, were conducted during a seven-week period in December 2015 and January 2016, with the researcher taking the role of an outside observer (Walsham 1995). Based on the participant's preference, 14 interviews were conducted in German and six in English. The interviewer was fully proficient in both languages.[25] This follows Myers and Newman's (2007) suggestion to create a friendly environment, noting the importance of interviewees being able to use their own language, which increases the likelihood of disclosure. Due to geographical and time limitations of some participants, two interviews were conducted via Skype, while the others were performed face-to-face in locations preferred by the participants. All interviews were recorded and then transcribed. Also, notes were taken during the interviews to capture complementary non-verbal insights.

Overall, we closely followed the principles proposed by Klein and Myers (1999) as well as those of Myers and Newman (2007). To ensure the coverage of all important questions, we conducted semi-structured interviews using a question script, which allowed for free development of the dialogue and assured a similar structure between all interviews. After all interview records were transcribed, the available data was iteratively analyzed using coding technique (Miles and Huberman 1984). Following this technique, we first applied open coding where the entire data was explored and broken apart to create codes. Subsequently, applying axial coding, we identified possible connections between codes and concepts (Corbin and Strauss 2008). This iterative process implied repeated examination of the interview data which gradually led to the elaboration of generalizations, i.e. factors and skills

[23] www.linkedin.com, www.xing.com, www.datasciencecentral.com

[24] www.meetup.com

[25] Quotations taken from the German interviews and used in the Results Chapter were faithfully translated into English language.

necessary for outsourcing data analysis tasks to online labor markets. Eventually, as no new insights were developed after 15 interviews, we concluded that data saturation has been reached and stopped interviewing after 20 interviews (Guest et al. 2006).

3.2 Quantitative research

After analyzing all interview transcripts, we developed an online questionnaire based on the results of the conducted interviews. To ensure traceability of results throughout the entire research project, and thus also support credibility, consistent term descriptions were used during all the research stages (Cronholm and Hjalmarsson 2011). Following a descriptive and cross-sectional research design (Pinsonneault and Kraemer 1993), the purpose of the questionnaire was to study the distribution of skills, expertise, and knowledge in the population of the most prominent freelance platforms. The cross-sectional design we adopted, implies that data was collected once and, thus, represents the population at that one point in time. We selected freelance platforms based on their size and on the availability of freelancers with data science or data analysis experience. After a thorough analysis of currently available freelance platforms, we chose Upwork and Freelancer as they constitute the biggest online workforce to date in general (~70%) and a large pool of freelancers specializing in different kinds of data analysis. For each platform we received 40 reliable survey submissions that have passed the quality assurance checks, making it a total of 80 (valid) participants, each of which was rewarded with 5 US dollars.

Throughout our survey we followed guidelines provided by Dillman (2011) and Fowler (2013). After the questionnaire was fully designed, it was iteratively pilot-tested with seven respondents. Short discussions with each person led to improvements and helped to refine the final version of the web survey. The final survey[26] consisted of 29 questions, spanning a mixture of primarily Likert-scale style questions as well as few open-ended, multiple-choice, and single-choice questions. The Likert-scale style questions were designated to capture the freelancers' skills, knowledge, and expertise with various tools, programming languages, and statistical methods. These questions were grouped into several matrices, had five-point response scales, and were all constructed in a similar manner (Dawes 2008). Other questions aimed at learning about the respondents' demographics, educational and occupational background, and experience with data analysis. In order to compare the opinions of interviewees and freelancers, we asked them which tasks of a data analysis project they could or could not imagine being outsourced to freelancers on OLM, and what difficulties they foresee in this undertaking. Since surveys, particularly online surveys, are subject to careless or inattentive responses (Meade and Craig 2012), we integrated several quality assurance questions to assure reliability of the collected data (Kittur et al. 2008; Gadiraju et al. 2015). As a result, 10 out of 90 submissions were excluded from the analysis. According to the findings of De Winter and Dodou (2010), the remaining 80 submissions were analyzed by

[26] https://form.jotformeu.com/60534618562356

performing one-sample two-tailed t-tests (which was found to perform comparably to the Mann-Whitney-Wilcoxon for the 5 point Likert scale) in order to identify statistically significant skills and expertise in tools and statistical methods. Moreover, for further analyses, we conducted descriptive statistics analysis and calculated Spearman rank sum correlations.

4. Results

In this section we first present the qualitative results of the interpretive case study (i.e. the interviews), and then outline the quantitative results of the web survey study.

4.1 Analysis of qualitative data

Our 20 interviewees comprised a diverse group of individuals that hold various positions and deal with data science or data analysis. Their experience varied between four and 22 years, with a median of 6.5 years. By having such a diverse range of professionals, we hoped to avoid elite bias, misrepresentation, and non-generalizable responses (Miles et al. 2013). The interviewees represent various industries, including insurance (4), financial service (4), digital analytics (5), analytics consulting (3), telecommunications provider (1), retail (2), and Internet broadcasting (1). The interviewees are all experts in data analysis and 95% of them stated that they have acquired their knowledge during university studies where 6 studied Economics, 3 studied Statistics, 3 studied Physics, and the rest studied other areas comprising mathematics and computer science classes. The majority, 60%, holds a master's degree, 25% hold a doctorate degree, and 15% have a bachelor's degree. Seventy percent state to have continued to learn on the job while 35% additionally have made use of online courses, books, or individual programming tasks to improve their skills.

a. Skills and knowledge required from Data Scientists (RQ1)

Although the main focus of the interviews was to extract the necessary skillset for freelancers, in order to identify communalities, we first asked interviewees to name the skills they need in their jobs. All interviewees declare technical skills to be absolutely necessary. These include **programming skills** (mentioned by 75% of the interviewees), **database skills** (55%), **machine learning skills** (30%), and general IT affinity like understanding how servers, software, and apps work (30%). Also **statistical and mathematical skills** are mentioned by almost every interviewee (90%), which includes the ability to conduct statistical analyses with various methods/tools and to have a general flair for numbers. P5[27] explained the necessity for statistical and mathematical skills as follows: '*Normally when we talk to stakeholders [...], they don't really agree or [...] understand why you need a big mathematics skill set, but I would argue [...] you still need a good mathematical background to make your conscious decision of the techniques you're using.*' Furthermore, **domain knowledge** (55%) in the field where the data analysis is conducted is considered to be of importance. It is necessary to understand the context and to know the company's business goals, as

[27] For the purpose of reporting we numbered the interviewees from P1 to P20.

this influences the direction of an analysis project. Interdisciplinarity is mentioned to be important in this context, as data scientists often have to tackle problems from diverse areas of a company. P14 stated: *'You need to have a pretty broad understanding, not always super deep, [...] to understand what you are actually analyzing.'* Another very important skill is **communication** (55%), i.e. the ability to communicate with different groups of interest and present results in a clear and simple way. If communication fails, the effort to analyze data can be in vain, as P3 stated: *'It doesn't help to just sit at the computer, you can be the best programmer, but at the end you need to be able to communicate the results clearly, and in the beginning you need to understand, what is important for the other person.'* Hand in hand with communication skills goes the flair for **consulting and mentoring** (25%), including working well with customers and keeping an open mind towards their needs. Having a **data-oriented mindset** (45%), i.e. to understand data and its structure, to be sensitive to the alternatives and limitations of data analysis, is another aspect interviewees mentioned several times. This skill of data understanding intersects with the understanding that data scientists need **experience** (35%) in their job in order to be successful (e.g. to know what is the best method to use in each case and how to interact with the data at hand). Moreover, data scientists need to have **logical thinking and reasoning** (45%), such that they are able to structure problems and to break them down, abstract, and operationalize solution steps. As stated, for instance, by P19: '*You need a structured approach, because if you're not a tiny bit structured, you start losing yourself in the data. So you should really always keep in mind, where do we want to go or which variable or function we want to optimize. And then anything we do, we target towards this goal. Otherwise we end up basically analyzing data which is not relevant for us.'* Oftentimes, data scientists are confronted with new problems in a new field, or have to apply new methods, tools, or software. Thus, they need to be **willing to learn** (20%) and to keep a **curious attitude** (20%). Curiosity is not only mentioned in connection with learning new things, but also as an aspect of being curious about what can be found in the available data. As stated by P16: '*This analytical curiosity, that you can simply recognize certain structures in the data itself, or also exercise patience to play around a little bit and to look in what direction the whole issue will develop*.' Also, data scientists and analysts need to be patient and enduring, as data analysis projects often require exhausting examination of data. P4 says: '*Sometimes you spend hours on some tools or some data and you don't find anything. So it's very hard sometimes to keep you motivated.*' Thinking about the problem deeply and trying to get to the bottom of it is therefore an important trait of data analysts. This includes paying attention to details and having an intuitive and inventive mind. This was mentioned by several interviewees (25%), e.g. by P7: *'You need to learn to pay attention to the last, to the smallest detail. You will, and that's guaranteed, stumble over those at the end.*' At the same time, they need to be skilled in **project management** (15%) as data analysis projects require adhering to constrained timelines. P10 highlighted: *'Oftentimes you work under time pressure. [...] You oftentimes also don't get the data that you need at the right time. So sometimes you have to prepare scripts blindly, then you receive the data and apply the scripts onto the raw data.'* Several interviewees (30%) note the fact that a data scientist cannot have all skills. Data science or analysis projects require **teamwork** and thus each analyst has a specific role with a corresponding skill set: *'Usually you just pick one or two max, that you try to perfect in a sense'* (P4). This skill set can be composed of the previously discussed skills. As to the software, a great majority of

the interviewees work with R (85%) and Python (75%) in their jobs. In general, the opinion prevails that one programming language is required but that it is irrelevant which one. Rather, it is important to '*having acquired all the programming logic in one or another language*' (P2). Also SQL (55%) and Excel (45%) are commonly used among the interviewees. Other tools mentioned several times include Java, JavaScript, Tableau, Hadoop-related technologies, ETL tools, and Oracle. Regarding statistical methods, P11 said: *'It's a full zoo of methods, over which you need to have a little overview.'* Mainly, interviewees utilize both descriptive statistics and data mining methods such as regressions, clustering, classifications, predictions and a number of machine learning algorithms. P16 looked at this topic with sorrow: *'90% of our jobs are descriptive. And all the cool stuff, that is really fun, is unfortunately done way too rarely.'*

b. Skills and knowledge required from freelancers (RQ1 continued)

Since we are interested in the entire skill set that should be present on a freelancing platform, we asked interviewees for the desired skills of freelancers for different tasks, and combined all answers to receive a full overview of necessary skills. **Statistical and mathematical skills** as well as **database skills** were mentioned most necessary by the interviewees (80% each). Next, **programming skills** (65%) are mentioned to be important. Specifically, almost every interviewee highlighted freelancers should know R (90%), followed by Python (65%). In general, similarly to the responses about their own skills, they regarded one programming language as necessary for freelancers without preferring one over the other (40%): *'I would say that it doesn't matter which programming language, because someone who knows one programming language, learns a new one very quickly'* (P13). Furthermore, freelancers should be familiar with SQL (45%), Excel (40%), ETL tools such as Talend (25%), and several other tools, data formats, and operating systems (mentioned by up to 20%). Also **domain knowledge** is mentioned by 65%. P20 explained this necessity as follows: *'Data analysis per se doesn't exist. It's always data analysis in a context: logistics, medicine etc. So the know-how in this context, in this domain, is absolutely necessary [...]. You need to explain first what the data means [...]. If people don't understand it, they cannot identify the data quality problems etc.'* **Data understanding**, the ability to understand data and its structure and how to work with it, was also identified as an important skill (50%). Forty-five percent mentioned **communication skills** that encompass being able to talk to different groups of interest, being an *'interface person'* as called by P7, and communicate results in a clear and straightforward way. P3 stresses the importance of communication: *'I think when it's external, it's even more important that there is good communication, because the person doesn't know the company well and the company doesn't know the person well. That's why you have to pay attention to communication all the more.'* Furthermore important are **visualization skills** (40%) (i.e. the ability to visualize data in a meaningful way), having an eye for simple and clear design, and attention to details, e.g., making sure *'that you don't make it red and green, but maybe orange and blue, because people have a red-green deficiency'* (P14). Having **experience** (30%) is an advantage, as it helps to make the right decisions, e.g., about the appropriate method. **Machine learning, text mining, documentation, and reporting skills** are also mentioned as necessary skills for freelancers by 10 to 20%. Moreover, some interviewees (25%) note that freelancers should have an **algorithmic and logical way of thinking**, which includes breaking problems down into smaller parts and deduct, as well as maintain a big picture view throughout their work.

Further, freelancers should be trustworthy, accurate, reliable, thorough, patient, and willing to learn new things. As such, those skills mirror the skills the interviewees ascribe to themselves.

c. Difficulties with outsourcing data analysis (RQ3)

Outsourcing data analysis tasks to freelancers is not necessarily an easy endeavor. Almost every interviewee is concerned about **communication issues** (80%). To begin with, a common language and shared understanding of the matter is necessary, which implies very clear requirements and well defined tasks. Since knowledge is sometimes assumed to be implicit and thus is not explicitly communicated between the freelancer and the customer, it can result in misunderstandings, inefficiencies, and in-transparency. Transparency (i.e., understanding of how the freelancer came to the results of data analysis) is crucial to guarantee the auditability of data (and the results). In addition, due to the distance, communication can take longer as one cannot simply go over to a colleague's desk but rather has to wait for a response (e.g., by email). Also, cultural differences may lead to communication problems, as noted by two interviewees. P2 noted: *'When you have foreign-language or diverse cultures, that maybe all speak English, but that maybe have a completely different understanding of a task, how to do it, and I would say the further away the cultures are from each other the more difficult it is.'* The fact that information can be lost during intermediate steps of communication, called *"Chinese Whisper Effect"* by P8, is another possibly occurring problem. Additionally, the initial **briefing** is a related issue (30%): *'That's the tricky part, getting the briefing right'* (P4). Freelancers need to know exactly what they have to do, so the briefing has to be very precise, which means additional time and cost. This in turn raises the question if it is worth to outsource: *'Oftentimes you ask yourself, should I rather do it myself or train somebody locally, after all it's an investment'* (P14). **High setup costs and time**, not only regarding the briefing, but also the infrastructure, is mentioned by 15% as a barrier to outsourcing data analysis tasks to freelancers, and thus, the effort has to be *'justified'*, as stressed by P3. Connected to this issue is also the **knowledge gap** (40%) that results from the complex IT environment of a company and the entire domain knowledge that freelancers first have to familiarize with. Another big issue is **privacy** and confidentiality of data (55%). Similarly to other respondents, P8 says: *'The problem with outsourcing is always that you actually don't want to give away the data, primarily for us they are customer data.'* This is also one of the reasons why some companies have not utilized outsourcing services so far and would feel uncomfortable sharing their data with freelancers. Ways to deal with this problem is anonymizing data or signing non-disclosure agreements. However, this problem still remains a sensitive issue. P6 points out the difficulty of finding a trade-off when he has to *"anonymize the sensitive data but to retain the utility of the data."* Furthermore, even if data is anonymized, it can happen that conclusions can be drawn about the actual identity of persons. Monitoring and **quality control** of work is another difficulty (40%). P5 says: *'If you were doing crowdsourcing, you have no guarantee, whatsoever, on the quality of the code or the piece of analytics that you get. So [...] somebody has to go and verify afterwards. And then it remains to be assessed [...], how much you benefit from crowdsourcing if you need to check afterwards what happened.'* **Trust** into freelancers (20%), the **vulnerability of data** whenever it is passed around (20%), the meeting of **deadlines** (15%), and the **danger of data manipulation** (15%), be it intentional or accidental, are other issues that have been

mentioned several times. P20 asks himself: *'Does the guy have a huge incentive to disappear with the data and supply the competition with the analyses? Can I prosecute him? Can I find him at all?'* P13 sees the following danger with data analysis in general and even more with freelancers: *'I can always steer data analysis a little bit in a certain direction. So if you pursue some interests and know some statistics, I'm not saying cheating, but bending statistics in a way that they claim something even if it's not really true.'* Thus, safeguards need to be applied. A problem that is not directly concerned with freelancers but rather with the company that is outsourcing data analysis is mentioned by 25% of interviewees: *'I think it's a huge problem that companies often don't even know what they can do in the first place'* (P3). Thus, they do not understand how they can draw meaningful insights through data science and hence have to be guided in the initial phase of exploring the possibilities of data analysis.

4.2 Analysis of quantitative data (RQ2)

We received 80 survey submissions from freelancers on Upwork and Freelancer, two most prominent OLM platforms. Respondents' experience with data analysis ranged from under a year to 45 years, with a median of 4 years. Noteworthy, many of the participants were beginners (i.e. freelancers with experience of just one year). Since data science is an emerging field, this could indicate that plenty of individuals are interested to start pursuing this profession. Another explanation can be the preference of beginners to seek for projects on OLM rather than through other recruitment channels. Lastly, we can not dismiss the option that reimbursement of 5 USD generated bias towards unexperienced freelancers, even though many freelancers stated that the reimbursement does not play a role in their decision to contribute to this study. Participants rated their expertise on average with 3.82 and median of 4 on a five-point Likert scale. They stated to have learned about data analysis on the job (22%), through university courses (22%), through the Internet (17%), books (16%), online courses (15%), and teaching videos (8%). Most freelancers, 76%, were male, whereas 24% were female. 51% were between the age of 25 and 34, 24% between 18 and 24, and 13% between 35 and 44; the remaining 12% spread between 45 and over 65 years. Almost half of participants were living in European countries, a quarter in Asia and the remaining in America, Africa, or Australia. All participants either had a university degree or were enrolled as students. The level of education was, thus, relatively high, with 28% holding a doctorate, 40% a master's degree, and 25% a bachelor's degree. Their field of studies encompassed mainly Computer Science, Mathematics and Statistics, as well as Engineering. The majority, 59%, were employed in full or part-time jobs, 19% were currently looking for jobs, and 18% were students. Due to the high employment rate, 45% spent only less than 10 hours per week on freelance work. Furthermore, 19% spent up to 20 hours, 15% up to 30 hours, 6% up to 40 hours, and 15% more than that. Since our interviewees mentioned that freelancers should have expertise in the field the data analysis is conducted, we asked freelancers in what domains they were experienced. As a result, Mathematics and Statistics were chosen as the most common domains, followed by Science and Research, IT, Engineering, Business and Management, Economics, and Finance.

In order to test all skills from the survey for statistical significance we performed t-tests by comparing the mean values of the responses to the test value 3, which

corresponds to having skills to a medium extent. Those items whose mean values were significantly larger or smaller than 3 were regarded as present, respectively absent on the platforms. Results show that all general skills, i.e. statistical, mathematical, database, programming, communication, presentation, visualization, machine learning, text mining, documentation and report writing skills, as well as data understanding were significantly different from 3. Thus, all mean values are larger, indicating that freelancers perceive their level of skills to be rather high, ranging from a mean of 3.28 in machine learning to a mean of 4.45 in data understanding (Table 6).

Performing the same tests for programming languages, tools, and data formats, we could obtain the following insights: Almost every item is significantly lower than 3. This implies that a lot of tools, programming languages, and data formats are not widely known by the surveyed freelancers. R, Python, and SQL, three items that were mentioned the most by our interviewees to be important for freelancers to have (90%, 65%, and 45% respectively), were found to be not significantly different from 3. This suggests on the one hand that these three tools are slightly better known than other tools, which had means lower than 3. But on the other hand, it also indicates that know-how for these tools is not highly present on freelancer platforms. Since particularly R and Python are some of the core skills required for data scientists (Kurgan and Musilek 2006), our findings support the fact that the widely discussed shortage of data scientists is also present on OLM (Harris et al. 2013). The only items having a statistically significant mean greater than 3 are Excel, PowerPoint, and CSV. Excel, with a mean of 4.21, is widely known among freelancers and they feel highly proficient in it. This is an important insight, as 40% of our interviewees stated Excel as a necessary skill to be known by freelancers. Again, we performed the same t-tests, this time for freelancers' knowledge of statistical methods. Even if freelancers are not very knowledgeable in statistical tools as indicated by the previous test, they state to have a high level of skills in statistical methods, particularly in descriptive and most inferential statistical methods. This is implied by the fact that most tested statistical methods have a mean value that is statistically significantly greater than 3. Machine learning techniques on the other hand have mean values lower than 3, partly statistically significant and partly not. This shows that these methods are not yet widely adopted by freelancers.

Table 6: T-test and Descriptive Statistics for General Skills

Variable	N	T-test			Descriptive Statistics				
		t-value	p-value	mean $\neq 3$	min.	max.	mean	median	std. dev.
Data understanding	80	21.116	.000***	Yes	3	5	4.45	5	.614
Communication skills	80	14.367	.000***	Yes	2	5	4.23	4	.763
Documentation/Report skills	80	11.075	.000***	Yes	2	5	4.08	4	.868
Presentation skills	80	9.786	.000***	Yes	1	5	4.00	4	.914
Visualization skills	80	9.649	.000***	Yes	2	5	3.96	4	.892
Mathematical skills	80	9.494	.000***	Yes	1	5	3.91	4	.860
Statistical skills	80	7.980	.000***	Yes	1	5	3.83	4	.925
Programming skills	80	6.749	.000***	Yes	1	5	3.80	4	1.060
Database skills	80	3.717	.000***	Yes	1	5	3.40	3	.963

Text Mining skills	80	2.895	.005*	Yes	1	5	3.34	4	1.043
Machine Learning skills	80	2.085	.040**	Yes	1	5	3.28	3	1.180

Legend: ***: 0.001, **: 0.01, *: 0.05 significance level

To test whether correlation exists between any surveyed items, such as freelancers' experience, skills, and expertise in various tools, programming languages, and statistical methods, we performed Spearman rank sum correlations. As expected, the more years of experience freelancers have, the higher they rank their level of expertise in data analysis (ρ=0.603**). In turn, with increasing level of expertise they rate almost every general skill significantly higher (except database, programming, and machine learning skills), and almost all inferential statistical methods (excluding most of machine learning techniques). Interestingly, freelancers proficient in Python are also proficient in Machine Learning (ρ=0.51), while those proficient in R are skilled in inferential statistics ((ρ=0.537**). Even though these two topics overlap to some extent, this distinction can indicate the division between the tools traditionally used by ML experts and statisticians.

To compare interviewees' and freelancers' opinions on the obstacles that exist when outsourcing data analysis, we also asked freelancers open-ended questions about possible hurdles. One difficulty they see is that the problem has to be very well defined and requirements and specifications have to be clearly set (40%). Also, it has to be ensured that freelancers understand the problem and the goal of the analysis project (8.75%). Otherwise, as mentioned by one respondent, the following issue could arise: *'Freelancers might misunderstand the main objective of the project, thus building different models or using less-satisfactory techniques to solve the problem at hand.'* In this regard, briefing (23.75%) is mentioned as an essential and also difficult phase of the outsourcing process, in which *'providing maximum information regarding the problem to be solved'* is necessary. Accordingly, communication is also mentioned by many participants (36.25%) as a difficulty when outsourcing data analysis to online freelancers. Knowledge gaps (13.75%) are also often identified obstacles in the outsourcing process, as e.g. mentioned by one respondent: *'Freelancers might not have the knowledge in the specific domain to conduct any meaningful interpretation of the results.'* This is why it is even more important to find and choose freelancers that possess the necessary skills for a given project. They also need to be reliable (21.25%). Furthermore, quality of work is seen as a problem and, thus, appropriate monitoring and control needs to be applied (16.25%). Interestingly, confidentiality of data is stated as a problem only by 6 participants (7.5%), indicating that freelancers are not aware of this problem. Time zone differences, language barriers, and providing an accurate scope regarding time and price are seen as hurdles each by five participants.

5. Discussion

Together, the qualitative and quantitative study results provide comprehensive information both about expected skills from freelancing data analysts and about the talent existing on major freelance platforms. Moreover, the interview results contribute to a better understanding of the obstacles towards outsourcing entire or parts of data analysis projects to OLM. Interestingly, the skills identified by the interviewed data scientists are not only limited to concrete skills picked up throughout studies such as math or coding (e.g. Kurgan and Musilek 2006), but go

much beyond and include various skills required for data analysis. In the following, we discuss the answers to our previously stated research questions.

(RQ1) The most prominent skills data scientists should have, in accordance with literature, are *mathematical/statistical skills* and *technical affinity* such as database and programming capabilities. However, *in addition to those, our interviewees emphasized the importance of domain knowledge and communication skills for the success of an outsourced analysis project*. Also having an *eye for aesthetics* and details when visualizing data is a trait that is not necessarily associated with data science, but was mentioned by many of the interviewees as very important. Moreover, possessing the above mentioned skills does not immediately lead to being a good data scientist but, rather, *a combination of hard and soft skills*, understanding of data and knowing how to get the most out of it, altogether represent a good data scientist. This includes understanding the limits of what can be achieved with the data at hand and the ability to communicate those limitations to the clients. They in turn, according to the interviewed data scientists, do not have a thorough understanding of data analysis and see data science as an oracle, capable of answering any kind of questions. These excessively high expectations could be attributed to the spread of data science buzz to the mainstream of decision makers in recent years.

(RQ2) All skills that interviewees expected freelancers to have were statistically significant when tested, with means ranging from 3.28 to 4.45 on a 5-point Likert scale (Table 7). Therefore, concluding from the data, the *necessary skills to perform data analysis projects exist on freelance platforms and outsourcing them, or parts of them, to online freelancers is a feasible task with regard to the skills*. Table 7 is arranged in decreasing order of freelancers' self-reported skills (last column in the table). Interestingly, the most highly ranked skills are abilities attributed to the general data understanding, communication, and documentation - skills that are similar to so called 'soft skills.' This can be explained with the subjective character of these skills and might hint to the need to find additional approaches to evaluate these abilities. On the other hand, freelancers feel most unconfident about skills such as text mining and machine learning. This can be explained with long specialization required in order to be proficient in these topics. We also asked data scientists what skills they have in order to ascertain whether they project their own set of skills to those required from online freelancers or seek for experts with complementary skills (first column in the table). The *skills that data scientists expected from freelancers much more than they had themselves were documentation, visualization, and database skills*. Conversely, data scientists did not expect freelancers to be as good as they are in advanced knowledge that requires mathematics, statistics, and machine learning. It seems like *data scientists might be interested in workers with a complementary set of skills that could perform tasks which do not require advanced knowledge but rather skills that allow to perform general tasks such as extracting data or preprocessing*.

(RQ3) The success of outsourced data analysis projects is not only determined by the availability of required skills, but several other factors. Particularly, both interviewees and freelancers saw *communication issues as the biggest hurdle when outsourcing data analysis*. This includes the necessity to have clear requirements about the project, conducting precise briefings with the freelancer, establishing shared understanding of tasks, and maintaining good communication throughout the project. Also *quality assurance and knowledge gaps* are aspects that were mentioned by

both interviewees and freelancers as hurdles in an outsourced project, whereas the latter laid even more emphasis on finding freelancers with appropriate knowledge and skills. *Privacy and confidentiality of data*, however, were mostly a concern of the interviewees; not as much by the surveyed freelancers. Hence, although outsourcing entails the hope to save resources in terms of time, money, and employees, outsourcing the project could require additional effort in terms of high setup costs and loss of time through additional communication, briefing, and performing quality assurance checks.

Table 7: Freelancer skills required according to data scientists compared with existing skills on freelance platforms

Skills…	**…Data scientists have**	**…Data scientists think that freelancers should have**	**…Freelancers feel they have**	
			Mean (1-5)	**Std. Dev.**
Data understanding	45%	50%	4.45***	0.614
Communication	55%	45%	4.23***	0.763
Documentation and writing	-	10%	4.08***	0.868
Presentation	15%	10%	4.00***	0.914
Visualization	10%	40%	3.96***	0.892
Mathematics	90%	80%	3.91***	0.86
Statistical skills	90%	80%	3.83***	0.925
Programming	75%	65%	3.80***	1.06
Database	55%	80%	3.40***	0.963
Text Mining	-	15%	3.34**	1.043
Machine Learning	30%	20%	3.28*	1.18
Domain Knowledge	55%	65%	Was asked directly	
Experience	35%	30%	Was inferred from years of experience and self-reported expertise on a 5-point scale	

Legend: ***: 0.1%, **: 1%, *: 5% significance

Furthermore, interviewees and surveyed freelancers stated *preprocessing data as the most suitable task to outsource to online freelancers*. However, it is also the task that entails significant difficulties when outsourced. Additionally, data scientists noted data preprocessing to be the most tedious task and the one that they would like to outsource. Despite the various obstacles in outsourcing this step they would be eager to see solutions that would allow to overcome these obstacles and therefore reduce their workload. *Problems identified during the interviews for the data cleaning process* were: (1) possibly confidential data, (2) the necessity of domain knowledge to understand the data, (3) the fact that data cleaning entails many tacit, subjective assumptions, and (4) the necessity of putting significant trust into freelancers when handing them data. Moreover, (5) data cleaning is a complex process, which requires a lot of customer contact and has to be repeated iteratively. Hence, the (6)

coordinating effort with the freelancer is significant, as it has to be constantly maintained, and (7) specifications and results have to be clearly conveyed. The next most mentioned task suitable for outsourcing to online freelancers, as mentioned by the interviewees, is *data collection*. Difficulties that are likely to arise are due to data spread over various sources such that freelancers first need to gain an understanding where the needed data is stored and how to get access to it. Also, data collection is error-prone and needs to be auditable. Additionally, several interviewees preferred to *outsource only the entire data analysis process*, since all steps within the project are connected: breaking them apart would lead to loss of knowledge and complicate the project. Problem definition and project specification were argued to be difficult to outsource by both interviewees and freelancers, as they require a lot of knowledge about the company, the domain, and the data. Interestingly, opinions about outsourcing data modeling were divided equally among interviewees. Again, knowledge about domain, background, and data were mentioned as obstacles in addition to the need to be able to handle large data, which require a lot of storage and computer power.

6. Limitations and Future Research

It is important to note that our approach has the following limitations. First, we have conducted an exploratory study, mainly bounded to the explicit factors related to the needs-supplies conceptualization. However, behavioral factors such as cognitive abilities, personal desires or satisfaction, and other psychological needs were not considered. Hence, future work will have to examine topics related to the behavioral specifications of candidates. Second, this study's cross-sectional design could be expanded to a longitudinal design to explore how abilities, skills, and expertise develop over time. Third, our study relies on the possible biased self-reported skills by the freelancers. Alternatively, one might want to assess the freelancers' abilities through tests. Whilst tests would be a more reliable approach to assess a *given* list of skills our design aimed at collecting comprehensive data from a big sample of freelancers active on different major OLM platforms to elicit an initial overview of the skills. Future research might sacrifice breadth and generalizability over different platforms for the sake of an unbiased assessment. Lastly, the geographical proximity of interviewees and their residence in Germany and Switzerland could be a potential argument against generalizability. We tried to address this limitation by interviewing a relatively large number of experts and ensuring that most of them work in international companies. Future studies will, of course, have to confirm the absence of a locality bias.

7. Conclusion

Most organizations experience an alarming shortage of data analysis experts in an emerging world of omnipresent data. To our knowledge *this study represents the first methodological attempt to explore the potential of overcoming the shortage of data analysis experts by outsourcing data analysis, or parts of it, to freelancers available on OLMs*. Specifically, we first explored the skills required for data analysis in general and then elicited the skills expected from freelancers on OLM from data scientists. We further investigated the obstacles to data analysis outsourcing and possible remedies to overcome them. The results presented here can be useful to 1) better understand the potential of OLMs as a remedy for shortage in data analysts, 2) for future

theorization on employee selection in online settings, and 3) serve as a starting point for extending the scope from data analysis to other areas common on online labor markets. The results of this study demonstrate that *the skills required for data analysis exist on major freelance platforms and that outsourcing data science projects, or parts of them, to OLMs is feasible*. These skills include data understanding, communication, documentation, presentation, visualization, and mathematical/statistical and technical abilities like programming, database, text mining, and machine learning. Furthermore, although data analysis outsourcing faces various hurdles (e.g., communication issues, knowledge gaps, quality of work, or data confidentiality), this study provides evidence that they can be resolved, thus, making outsourcing data analysis tasks possible. As such, it highlights a possible approach for overcoming the scarcity of data science professionals.

8. Acknowledgments

We would like to sincerely thank the interviewees and the freelancers that took part in this study. We are also thankful to the anonymous reviewers for their efforts in reviewing this paper. This work was supported in part by the Swiss National Science Foundation (SNSF- Project: 200021- 143411/1).

9. References

Agrawal, A, Horton, J, Lacetera, N and Lyons, E, 2013 *Digitization and the contract labor market: A research agenda.* National Bureau of Economic Research.

Alam, S and Campbell, J 2014 *Examining cultural volunteer crowdsourcing technology: An appropriation perspective.* In 35th International Conference on Information Systems (ICIS 2014), Auckland, New Zealand.

Bernstein, A, Klein, M, and Malone, T W 2012 *Programming the global brain.* Communications of the ACM, (55:5), pp.41–43.

Chatfield, A T, Shlemoon, V N, Redublado, W, and Rahman, F 2014 *Data scientists as game changers in big data environments.* ACIS.

Christoforaki, M and Ipeirotis, P G 2015 *A system for scalable and reliable technical-skill testing in online labor markets.* Computer Networks, (90), pp.110–120.

Corbin, J and Strauss, A 2008 *Basics of qualitative research*. Thousand Oaks, CA: SAGE, 3rd edition.

Cronholm, S and Hjalmarsson, A 2011 *Experiences from sequential use of mixed methods.* The Electronic Journal of Business Research Methods, (9:2), pp.87–95.

Crowdflower 2015 Crowdflower 2015 *data scientist report.* Technical report.

Davenport, T H and Patil, D 2012 Data scientist: *The sexiest job of the 21st century.* Harvard business review, (90), pp.70–76.

Dawes, J G *2008 Do data characteristics change according to the number of scale points used? An experiment using 5 point, 7 point and 10 point scales.* International journal of market research, (51:1).

De Winter, J C and Dodou, D 2010 *Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon. Practical Assessment, Research and Evaluation*, (15:11), pp.1–12.

Denzin, N K 1973 *The research act: A theoretical introduction to sociological methods*. Transaction publishers.

Dillman, D A 2011 *Mail and Internet surveys: The tailored design method–2007 Update with new Internet, visual, and mixed-mode guide*. John Wiley & Sons.

Feldman, M and Bernstein, A 2014 *Cognition-based task routing: Towards highly-effective task-assignments in crowdsourcing settings.* In 35th International Conference on Information Systems (ICIS 2014), Auckland, New Zealand.

Fowler, F J 2013 *Survey research methods*. Sage publications.

Gadiraju, U, Kawase, R, Dietze, S, and Demartini, G 2015 *Understanding malicious behavior in crowdsourcing platforms: The case of online surveys.* In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp.1631–1640. ACM.

Guest, G, Bunce, A, and Johnson, L 2006 *How many interviews are enough? An experiment with data saturation and variability.* Field methods, (18:1), pp.59–82.

Haas, D, Ansel, J, Gu, L, and Marcus, A 2015 *Argonaut: macro-task crowdsourcing for complex data processing.* Proceedings of the VLDB Endowment, (8:12), pp.1642–1653.

Harris, J G, Shetterley, N, Alter, A E, and Schnell, K 2013 *The team solution to the data scientist shortage.* Accenture Institute for High Performance.

Howe, J 2008 *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House.

Ipeirotis, P G 2010 *Analyzing the amazon mechanical turk marketplace.* XRDS: Crossroads, The ACM Magazine for Students, (17:2), pp.16–21.

Johnson, R B and Onwuegbuzie, A J 2004 *Mixed methods research: A research paradigm whose time has come.* Educational researcher, (33:7), pp.14–26.

Kamar, E, Hacker, S, and Horvitz, E 2012 *Combining human and machine intelligence in large-scale crowdsourcing.* In Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, (1), pp.467–474. International Foundation for Autonomous Agents and Multiagent Systems.

Kandel, S, Paepcke, A, Hellerstein, J M, and Heer, J 2012 *Enterprise data analysis and visualization: An interview study*. IEEE Transactions on Visualization and Computer Graphics, (18:12), pp.2917–2926.

Kittur, A, Chi, E H, and Suh, B 2008 *Crowdsourcing user studies with mechanical turk.* In Proceedings of the SIGCHI conference on human factors in computing systems, pp.453–456. ACM.

Klein, H K and Myers, M D 1999 *A set of principles for conducting and evaluating interpretive field studies in information systems.* MIS quarterly, (23:1), pp.67–93.

Kokkodis, M and Ipeirotis, P G 2015 *Reputation transferability in online labor markets.* Management Science, *Articles in Advance*, pp.1-20.

Kurgan, L A and Musilek, P 2006 *A survey of knowledge discovery and data mining process models.* The Knowledge Engineering Review, (21:01), pp.1–24.

Meade, A W and Craig, S B 2012 *Identifying careless responses in survey data.* Psychological methods, (17:3), pp.1-20.

Miles, M B and Huberman, A M 1984 *Qualitative data analysis: A sourcebook of new methods*. JSTOR.

Miles, M B, Huberman, A M, and Saldana, J 2013 *Qualitative data analysis: A methods sourcebook.* SAGE Publications, Incorporated.

Mill, R 2011 *Hiring and learning in online global labor markets.* Stanford University, Stanford, CA.

Myers, M D and Newman, M 2007 *The qualitative interview in is research: Examining the craft*. Information and organization, (17:1), pp.2–26.

Pallais, A 2013 *Inefficient hiring in entry-level labor markets.* Harvard University and NBER.

Pinsonneault, A and Kraemer, K 1993 *Survey research methodology in management information systems: an assessment.* Journal of management information systems, (10:2), pp.75–105.

Serban, F, Vanschoren, J, Kietz, JU and Bernstein, A, 2013 *A survey of intelligent assistants for data analysis.* ACM Computing Surveys (CSUR),*45*(3), p.31.

Suzuki, R, Salehi, N, Lam, M S, Marroquin, J C, and Bernstein, M S 2016 *Atelier: Repurposing expert crowdsourcing tasks as micro-internships.* In CHI 2016, San Jose, California, USA.

Teddlie, C and Tashakkori, A 2009 *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage Publications Inc.

Verroios, V, Papadimitriou, P, Johari, R, and Garcia-Molina, H 2015 *Client clustering for hiring modeling in work marketplaces.* In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.2187–2196. ACM.

Walsham, G 1995 Interpretive case studies in is research: nature and method. *European Journal of information systems*, (4:2), pp.74–81.

Walsham, G 2006 *Doing interpretive research*. European journal of information systems, (15:3), pp.320–330.

Zikopoulos, P C, Eaton, C, DeRoos, D, Deutsch, T, and Lapis, G 2012 *Understanding big data*. New York: McGraw-Hill.

Towards Collaborative Data Analysis with Diverse Crowds – a design science approach

This chapter is based on a paper accepted in **Conference on Design Science Research in Information Systems and Technology (DESRIST)**

We presented a short version of it at the Collective Intelligence conference 2016 as a poster (Feldman et al. 2016).

Towards Collaborative Data Analysis with Diverse Crowds – a design science approach

Abstract

The last years have witnessed an increasing shortage of data experts capable of analyzing the omnipresent data and producing meaningful insights. Furthermore, some data scientists mention data preprocessing to take up to 80% of the whole project time. This paper proposes a method for collaborative data analysis that involves a crowd without data analysis expertise. Orchestrated by an expert, the team of novices conducts data analysis through iterative refinement of results up to its successful completion. To evaluate the proposed method, we implemented a tool that supports collaborative data analysis for teams with mixed level of expertise. Our evaluation demonstrates that with proper guidance data analysis tasks, especially preprocessing, can be distributed and successfully accomplished by non-experts. Using the design science approach, iterative development also revealed some important features for the collaboration tool, such as support for dynamic development, code deliberation, and project journal. As such we pave the way for building tools that can leverage the crowd to address the shortage of data analysts.

1. Introduction

Data analysis is a complex task that touches on many skills. Experts conducting data analysis are, therefore, expected to be proficient not only in the domain of their interest, but also in other disciplines such as statistics, computing, software engineering, and algorithms (Davenport & Patil 2012). These high expectations make data scientist scarce, leaving their valuable services out-of-reach for a big share of public. This also means that the way to become data analysis expert is extremely complex and the specialization can not be easily gained.

In this paper, we introduce an approach for collaborative data analysis *to allow non-experts to cooperate on data analysis projects*. In contrast to the lack of data scientists, there are many freelancers or enthusiasts that have some basic coding skills obtained either in introductory classes during their studies or self-acquired throughout the course of their life. While those non-experts do not have all necessary skills to perform end-to-end data analysis projects, they can be involved in some parts where their skills are sufficient. Specifically, we argue that non-experts with some coding skills can be especially helpful in the *data preprocessing* stage of data analysis. In this step data scientist transforms raw data into a data suitable for statistical modelling, as it is often inconsistent, incomplete and contains many errors. It is, therefore, likely that prior to statistical modelling, which requires significant knowledge in statistics and computer science, there is a need in "data wrangling" – transforming and editing raw data until it is suitable for data analysis (Kandel et al. 2011).

At the same time, data preprocessing and the following statistical analysis can not be decoupled. Often, in order to apply certain statistical approaches, the data has to be previously transformed and organized accordingly. For instance, to apply a statistical model that assumes linearity the dependent variable often has to be transformed first. Moreover, data analysis is an iterative process where data

preprocessing and modelling are intertwined: the results of data analysis lead to new ideas on how better to analyze data, which in turn leads to additional data preprocessing. Therefore, it is important that experts and non-experts cooperate and efficiently coordinate tasks. Following these considerations, we propose a process where data analysis projects are divided into sub-tasks and each is assigned to a freelancer with limited knowledge in data analysis and (some basic) coding skills. While the participants are assigned to different tasks, they interact through various communication channels in order to draw on their collective knowledge (Bernstein et al., 2012), and thus, reach the desired results. Dividing the project into several simple tasks allows project manager – a data analysis expert responsible for the whole data analysis project – to distribute and coordinate the tasks. This way the manager can take advantage of various workers' abilities in order to conduct data analysis. In our experiments, we explore whether the results of non-expert teams orchestrated by a manager are comparable to the results produced by experts handling the whole project. Therefore, our goal is to propose a practical solution to the problem of shortage of data scientists and allow non-experts to take part in the process of data analysis.

Our contributions are as follows: First, we present a method for collaborative data analysis in online freelance setting. Second, through a set of experiments, we show that the proposed approach is both cost-effective and can produce results with equivalent quality to those produced by data scientists. Finally, following a design-science approach, we develop a platform that supports collaborative data analysis with mixed-level expertise.

2. Literature review

In the following section, we introduce prior work on which we based our study. Its subsections review the success factors of online collaboration, describe the existing solutions for collaborative data analysis, and discuss the theoretical underpinnings that informed our method.

Online Collaboration: The advance of communication technology as well as a spread of sociotechnical systems made it possible for workers effectively collaborate within a distributed environment. Rather than meeting face-to-face, workers can rely on various communication channels such as emails, teleconference software or chat tools to cooperate in various tasks (Sere et al. 2011). Many domains adopted computer mediated collaboration as a useful tool for reaching goals. Scientists use different online tools to engage in research discussions and activities (Van Noorden 2014). Educators take advantage of online collaborative learning techniques to support students in achieving competence and foster skills like team working and group decision making (MacDonald 2003). Moreover, online collaborative tools facilitate marketing and decision making activities by, for instance, allowing better understanding of shopping behavior and predicting demand for products (Yadav & Pavlou 2014). Previous research has identified multiple factors that impact successful online collaboration. First, a team has to be supported by senior member or manager who is facilitating the progress of the task and provides feedback (Tseng et al. 2009). Second, the members have to make themselves familiar with each other, which in turn should lower the psychological barrier of estrangement and promote cooperation over time (Salehi et al. 2016). Third, well-established communication is

essential to avoid disagreements about the priorities and strategy to achieve pre-set tasks (Yukl 2001). Fourth, trust along the group members supports the feeling that all members work towards the same goal and make every effort to achieve the best possible outcome in order to earn trust among team members. Finally, the last element is well established organization of the team. A competent leader will support the team in the process of developing manageable and effective workflow to accomplish the task in short time end with reasonable efforts (Tseng et al. 2009; Salehi et al. 2016). We considered all these factors during the design of the artifacts that will support collaborative data analysis with non-experts.

In crowdsourcing literature, a few notable methods to support crowd-collaboration have been proposed. For instance, Turkomatic is a tool that utilizes crowdworkers to plan and execute complex tasks. Requesters can watch workers decomposing and solving tasks in real time, either collaboratively or independently. Requesters can intervene to modify tasks or request new solutions to subtasks as needed (Kulkarni et al., 2012). Another framework, CrowdForge, introduces a map-reduce paradigm to split complex work into small parts and solve it in crowdsourcing setting. The task is broken into multiple subtasks that are concurrently solved and verified by other workers, and eventually merged into a cumulative output. However, although the framework relies on a powerful paradigm of parallel work execution, it assumes that complex work can be decomposed into lots of merely dependent micro tasks – an assumption that is often violated (Kittur et al. 2011). Other notable examples of online collaboration in crowdsourcing are CrowdWeaver – supporting with visual interface for real-time managing both human and machine crowdtasks within an integrated workflow (Kittur et al. 2012) and Soylent – a word processing interface, implementing the Find-Fix-Verify crowd programming pattern, which splits tasks into a series of generation and review stages and utilize the collaboration among crowdworkers through independent voting and agreement to produce reliable results (Bernstein et al. 2010).

Existing solutions for collaborative data analysis: One of the most well-known examples of collaborative data analysis is Kaggle (Carpenter 2011). Kaggle is a web platform for data analysis that allows organizations to post their data projects and invite enthusiasts all around the world to participate in contests. Participants experiment with different techniques and compete against each other to produce the best models. For most competitions, submissions are scored immediately, based on their predictive accuracy relative to a withhold test-set of data, and summarized on a live leader-board. Once the deadline is over, the competition host pays prize money in exchange for the winning model (Dissanayake et al., 2014). Participants are allowed to team up together to collaborate on projects, and thus improve their chances to win the contest. Other solutions, such as Sense.us (Heer et al., 2009) or Many Eyes (Viegas et al. 2007), have been proposed for collective data analysis by enabling crowds visually inspect data. For example, Willet et al. (2011) presented CommentSpace, a collaborative tool for visual analysis that allows to annotate graphic content with tags and links that reflect the relationship between comments and visualizations. Wisteria and Wrangler are example of two human-in-the-loop systems that involve crowds in data cleaning by inferring the operations performed manually by crowds and extrapolating them to the whole dataset (Haas et al. 2015; Kandel et al. 2011). Collaborative data analysis can be seen as an offshoot of

distributed software projects. However, despite the evolution of advanced collaboration and software engineering tools (e.g., GitHub, Jira), software development is still mostly a prerogative of experts and does not involve laymen.

All mentioned solutions fall short on supporting collaborative data analysis by relying on crowds with mixed expertise. While platforms such as Many Eyes or Wrangler appeal to crowds without any prior expertise, platforms like GitHub require substantial skills in order to be able efficiently collaborate using their functionalities. Moreover, web-portals for crowdsourced data science such as Kaggle or TopCoder are rather a meeting point for data scientists and customers and, by and large, do not support the teams with any functionalities throughout data analysis.

Theoretical underpinnings: Tasks can be complex and may involve the coordination of a large number of participants with different capabilities. Therefore, different scientific communities have made efforts to associate tasks by decomposing them into the sub-tasks required to complete the full task (Campbell & Wu 2011; dos Santos & Bazzan 2009). For instance, within the AI community, Chandrasekaran et al. (1992) proposed a hierarchical task-method decomposition, which recursively links a task to alternative methods and their subtasks. This method emphasizes modeling of domain knowledge by utilizing tasks and methods as mediating concepts and, therefore fits our scope of the data analysis domain. Stefik (1981) proposed an approach of constraint hierarchical planning, where the constraints are dynamically formulated and propagated as the process proceeds. Subsequently, these constraints are used to coordinate the solutions of defined sub-tasks. The organizational approach, as presented in the Handbook of Organizational Processes of Malone et al. (1999), in contrast, introduces methodologies to represent and codify organizational processes and provides different perspectives on how business processes might be decomposed into sub-activities. A difference between these two approaches lies in their different purposes: while AI is focused on building computer systems that automatically execute processes, the organizational approach advocates building systems to support people to plan and execute processes. Howison and Crowston (2013) propose a theory of collaboration through open superposition. Developed in the context of open source software development, this theory emphasizes that tasks that appear too large for individual are likely to be postponed until they redefined such that they can be performed by single member, and that most of the tasks are indeed accomplished with only a single programmer.

These theories inform our solution in a few ways. A) decomposition of ill-defined task has to be tied into domain knowledge. B) the envisioned system should enable experts to decompose the task in efficient manner (e.g. through taxonomy or by utilizing expert's knowledge). C) There is a need for efficient coordination and communication in order to enable unimpeded process of data analysis. D) data analysts working on a well defined task will prefer to work on their own rather than collaboratively in an online team. However, they will be interested to coordinate the outputs of their task, to discuss possible solutions, and to receive feedback to their job. E) every task assigned to a worker should be well adapted to the skills and needs of the worker, with a clear specifications and and task manager that can supervise and help with advice and guidance.

3. Research Design

The research design presented here follows a design-science research approach in information systems as presented by Hevner et al. (2004). The authors describe design-science process as a sequence of expert activities that produces a set of artefacts with the following evaluation and feedback in order to improve both the quality of the artefacts and the design process. According to the theory taxonomy proposed by Gregor (2006), the proposed research resides within the *theory for design and action* by contributing to knowledge via addressing the considerations of a) the utility to a community of users, b) the novelty of the artefact, and c) the persuasiveness of claims that it is effective. As the goal to define and develop artefact that supports a novel approach of *collaborative data analysis with mixed-expertise crowds*, design can be seen as a search process involving an iterative evaluation and refinement of artefacts (Hevner et al. 2004; Reinecke & Bernstein 2013). The research methodology we adopted follows Peffers et al. (2007) and includes six steps: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication (see also Figure 6).

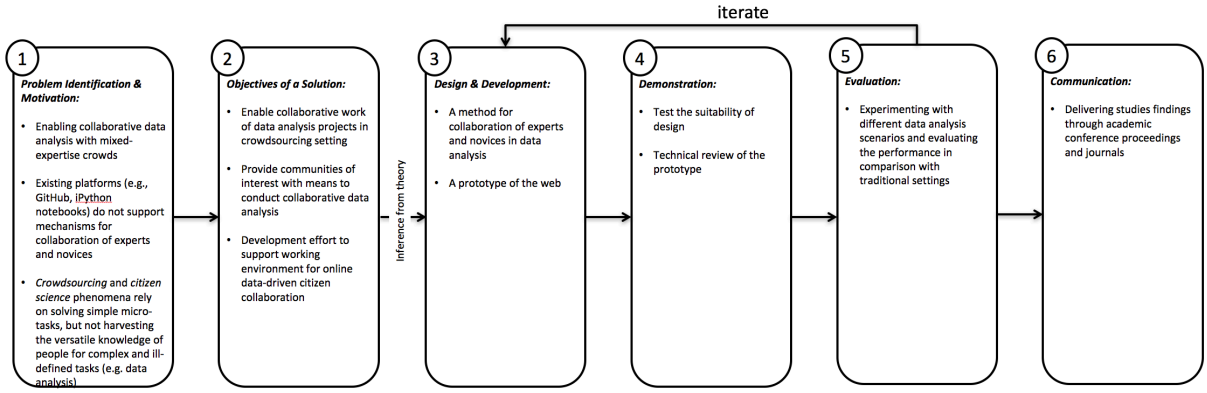

Figure 6. Research methodology (Following Peffers et al. 2007)

Following the figure, we start by laying out the the research motivation: (a) to enable collaborative data analysis by crowds with different expertise, (b) the lack of platforms that support an efficient environment for data analysis for non-experts in a dynamic manner, and (c) to leverage the crowdsourcing and citizen science phenomena of harvesting knowledge that is hard to reach. We then define objectives of the solution: (a) to enable collaboration on data analysis tasks on web, (b) to provide communities of interest with means to conduct collaborative data exploration, and (c) to propose a web environment for online collaboration. At the design stage, to the best of our knowledge, no dominant method has been identified so far to incorporate people with diverse skills into data analysis. Hence, the major challenge of this paper is defining and evaluating the needs for collaborative data analysis, accounting for the diverse nature of crowd workers. To do so, we start with the top-down approach of expert managing the novices and gradually explore the predominant factors for successful collaboration and tasks' coordination. The results will be demonstrated through the web application prototype built based on the discussed artefacts and set of experiments in which we evaluate the crowd's performance on a series of data analysis projects to check whether the designed prototype satisfies the prerequisites.

3.1 Conceptualization of the artefact (data analysis tool)

In this study we present a framework that allows non-experts to work on data analysis projects. Our framework i) supports a project manager in decomposing complex tasks into small and facile sub-tasks, ii) supports coordination and supervision by project manager, and iii) enables an iterative development of the data analysis project. The methodology we propose implies that the project manager defines a project and distributes assignments to workers in a top-down approach. A top-down approach is considered as more appropriate for well-specified, rather than ill-defined problems (Redmiles 2000). However, we decided in favour of this method, as the scenario we envision is of non-experts that are competent to perform preprocessing tasks only with the appropriate supervision. It is, therefore, necessary to impose task decomposition hierarchy to be able to manage the complexity of task on the expense of its flexibility. In addition, as our approach implies iterative exploring of the success factors for the scenario we investigate, the top-down approach is better suited for understanding how strictly hierarchical approach can transition into more collaborative one. For instance, it allows to see throughout the iterative development and evaluation, where the expert oversight can be replaced with peer-review of other novices, how decisions made throughout data analysis can be informed by the broad knowledge of the crowd to enrich expert's decisions, or how to establish effective communication to unleash the untapped knowledge of project members. Following the design science approach, we conducted two iterations of prototype development with consequent evaluations. In the following we first describe the general workflow and then the evolvement of the prototype and of the methodology after each iteration.

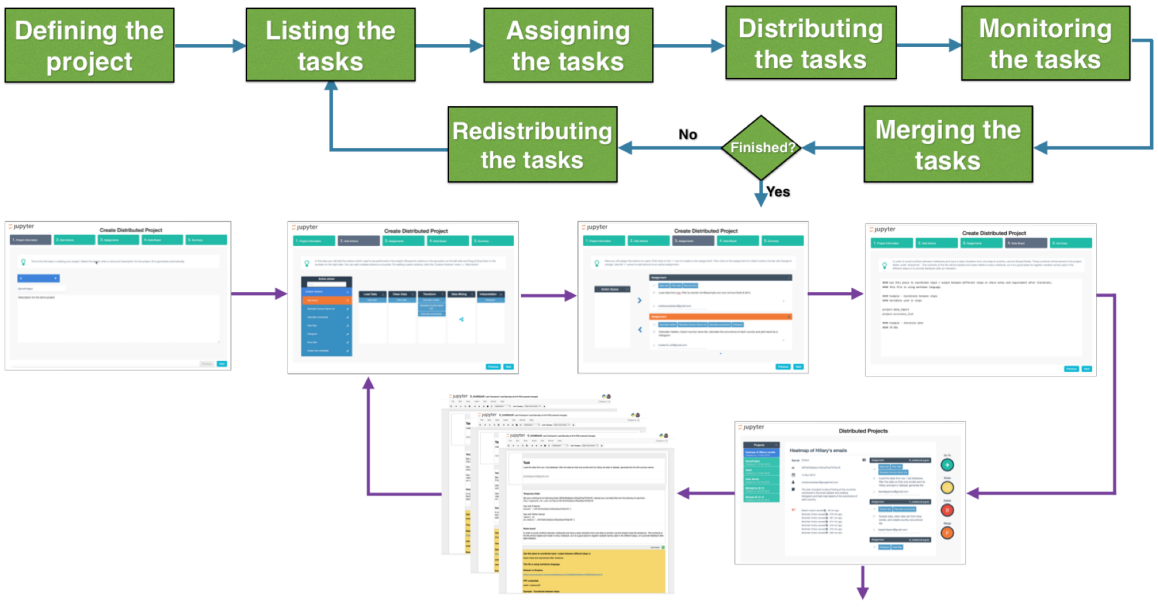

Figure 7. Process workflow

Figure 7 describes the workflow of the envisioned collaborative data analysis project. The figure presents both schematic workflow of the process on the top and the corresponding print-screens of the prototype on the bottom. The first part of the workflow is focused on the project definition, task decomposition and sub-tasks assignment processes done by the project manager. The second part focuses on the

iterative collaboration on the project, enabling the manager and team to refine the implementation and output through multiple iterations.

Next, we go through each of the workflow steps and explain them. First, the project manager defines the project by entering all relevant details, such as the software language to be used, the project name, and the project description. This step also provides some validations, ensuring that all necessary information is present. Next, the manager lists and defines the actions that need to be done. An action is the smallest unit of sub-task and an assignment is a composition of actions assigned to a worker. An example of an action would be *loadFromCSV*, which receives as input the path of the CSV file and returns a data-frame. Splitting assignments into small actions, especially in the preprocessing part, allows the project manager to distribute them to non-expert workers and supervise their execution throughout the assignment. Further, **tasks are assigned** to suitable workers. The assignment of tasks to workers follows a top-down approach and can be done on the basis of different criteria such as worker or task attributes, or by taking into consideration external factors such project deadlines or budget. The **tasks are then distributed** among the workers by virtue of email invitation to the IPython (or jupyter) Notebooks that are created and contain all the required information. At this point the workers can work on their personal notebooks stored on their personal cloud storage (Google Drive) and interact with each other through the shared notes-board. They can also review others notebooks and comment on the relevant code using side-comments. All throughout the project, manager can **monitor the progress** of the workers and guide them towards the desired output. Finally, the tasks are merged into one notebook, that allows a manager to run the end-to-end implementation. Project managers can then verify that the output meets their expectations and that the interaction between different assignments works properly. Otherwise, if the goal has not been reached, the implementation of the tasks will be changed or new **tasks will be redistributed** and the project will enter a new iteration.

3.2 Iterative Development-Demonstration-Evaluation

The proposed solution has been developed in two iterations by improving the method and the web-prototype for collaborative data analysis in a consecutive manner. Based on the evaluation of each iteration, we focused on advancing the artifact with respect to the following two criteria:

First, the proposed methodology and web-prototype should enable coordination and successful completion of data analysis projects with diverse crowds. Specifically, typical data analysis projects should be decomposed into subtasks such that they will be simple enough to be performed by non-experts. We evaluate these criteria qualitatively, through a user study by answering the following hypothesis:

> *H1: It is possible to decompose typical data analysis projects into small enough tasks such that the complexity of these tasks is substantially reduced.*

Second, the proposed solution has to be comparable in quality to traditional expert-based data science projects. To answer whether the proposed methodology is feasible and can reach the desired output of collaborative data analysis with mixed-level expertise teams, we propose the following hypothesis:

H2: The quality of the results produced by a team of non-non experts is comparable to the one achieved by experts.

In the following we will present three versions of the prototype and discuss their performance according to these measures. Note that we tested all iterations on real-world examples chosen from Kaggle based on the following criteria: a) the projects should be implemented either in R or Python, as these are the most popular languages in data analysis, b) the projects should contain a relatively large preprocessing part, as that has been found to be a major part of data (Krishnan et al. 2015), c) the projects should encompass various types of data analysis such as descriptive statistics, visualization, and prediction, d) the projects should be conducted by individuals that can be considered as experts, either based on their verified biography or because of their high ranking on Kaggle, and e) the projects should not be trivial (i.e., we limited the minimal size of the project to be about 150 lines of code, chose projects with significant number of up-votes, and history of comments such that it can be assumed that the code went through a substantial public review).

3.2.1 The pilot study

Following to literature review we designed the first prototype of our tool. The web-platform is based on the Jupyter Notebook (colloquially known as IPython notebook) and available online. Jupyter is a command shell for interactive computing in multiple programming languages that offers enhanced introspection, media, additional shell syntax, tab completion, and rich history. Using Jupyter, researchers can capture data-driven workflows that combine code, equations, text and visualizations and share them with others. We decided in favor of this platform due to the following reasons. First, it is a browser-based notebook with support for code, text, mathematical expressions, inline plots, and other rich media. These functionalities are essential for collaborative data analysis as they allow participants to exchange results and easily communicate their findings and difficulties. Second, although initially designed for Python, the platform is language agnostic and provides the ability to be extended with additional interpreters such as R and Ruby. Third, this platform supports an interactive data visualization toolkit, often required in data analysis.

To better understand the requirements of the proposed solution, *we conducted a user-study with three graduate students supervised by a PhD student.* As part of their course work, the students conducted data analysis project that involved substantial data preprocessing followed by network analysis. The supervisor was managing the task decomposition and divided the project among the group members with further coordination of the process up to its successful accomplishment (following the process presented in Figure 7).

The goal of pilot study was twofold. First we wanted to reach a proof-of-concept, showing that our approach is feasible and data analysis projects can be successfully accomplished with non-experts. Therefore, we alleviated some constrains such as performing the experiment in real-world setting using freelancers/crowdworkers or assuring that the analysis has been performed exclusively on our platform. Second, we aggregated the feedback to better understand the requirements of the proposed

tool and to evaluate the workflow. In addition, the feedback received from this iteration helped us to simplify the coordination process and to resolve some technical issues.

Conclusions/requirements drawn from pilot study and their addressing: I) All participants pointed to the need for collaboration and communication tools. While some can be externally used (e.g., forums, video chats), some tools have to be embedded into the platform to support effective coordination between team members. Especially, since the assignments distributed to workers are often interdependent, it is important to allow team members to comment on the relevant code-blocks of their peers. To address this need, *we developed features that allow workers better to collaborate*. For instance, we presented "sticky notes" – a note that every team member can leave next to the code-box of a Notebook. II) Another point, raised by the manager, is to improve the control over the project by enabling easy access to the notebooks, evaluating the current results, and (re)distributing the tasks. We, therefore, *added a functionality to automatically merge the notebooks into a master notebook that includes all notebooks in predefined order*. This allows to run all distributed assignments at one run and quickly identify bugs and inconsistencies. To redistribute the tasks with new instructions, we implemented a feedback loop (see Figure 7) that allows easily to redistribute the tasks to team members with new instructions and based on the previously submitted code. III) To improve the collaboration, team members pointed to the need to have access to the instructions every team member received from the manager as well as have the opportunity to intervene in order to clarify what in their opinion has to be done. To address this, *we added a project journal, where all project participants can add their comments*.

Note, while such functionalities exist in professional software development platforms such as GitHub, our goal is to enable *non-experts* to collaborate instantly on data analysis projects in easy and interactive way with no knowledge on the principles of distributed software development. In the following iterations, we qualitatively evaluated the proposed features and extend our platform according to the additional feedback provided by crowdworkers in the real-world setting. Most of the attention in the following two iterations though, is devoted to testing the postulated hypotheses.

3.2.2 First iteration - three data analysis projects

For a real-world evaluation, we selected three data analysis projects that represent various types of data analysis. Projects were taken from a large crowdsourcing data science platform, Kaggle. In these experiments, a data analysis expert (also a co-author) assumed the role of the project manager and the workers are recruited through the Upwork[28] platform. As of today, Upwork is the biggest online labor market and contains online freelancers in different domains. Data analysis is one of its most common domains and has a large pool of freelancers with different level of expertise willing to work on data analysis projects (Agrawal et al. 2013). These tasks can be classified as of moderate complexity as they involved mainly data

[28] www.upwork.com/

preprocessing and visualization, and did not require any advanced knowledge in data analysis.

Task #1: Earnings Chart by Occupation and Sex[29]*:* The aim of the first project is to create a chart showing the earnings of the population by occupation and gender, using the data of the latest US census from 2014. The original Kaggle project analyzes 24 occupation categories, while in our project we randomly selected 11 categories. The workers had to classify the list of the professions into these 11 job categories (e.g., management, science, military) and plot a chart of the earnings for each occupation with respect to the gender. This project is the easiest and was accomplished in two days.

We split the project into three assignments. The first assignment involved data loading and cleaning with the primary goal of identifying the correct industry code ranges and sub-setting the data. It consisted of five actions. The first was to *Identify Occupation Industry Codes*, and *Subset data* and the output of this task was a file containing the information about the population working in the 11 industries relevant for our chart. The second task focused on the data transformation and had only two actions – *Mean* and *Save results*. The output of this task was an aggregated data set containing the mean earnings of men and women per industry. In the last task, the crowdworker had to plot the data as a bar chart diagram in descending order, showing the distribution of men and women per industry and their average earnings. It consists only of one action – *Bar Chart*, and produced as output a bar-chart similar to the one in the Kaggle project.

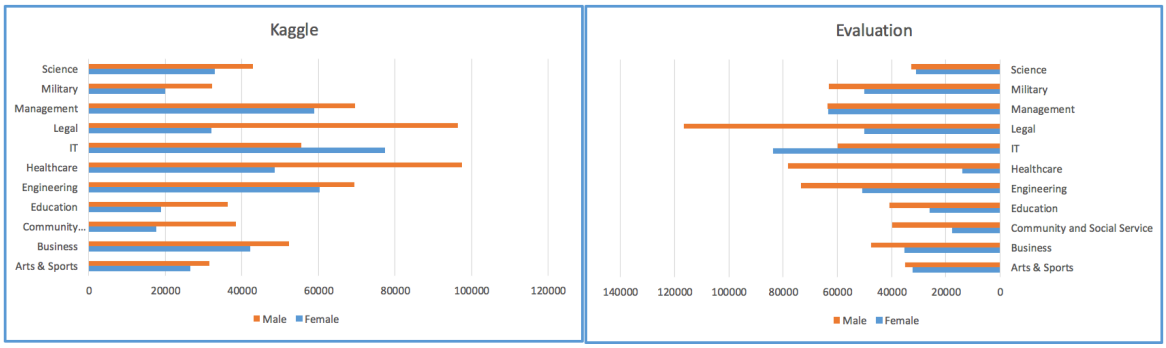

Figure 8 Pearson correlation coefficient ϱ=0.8

The main focus of this project was to find the right occupation categories and to subset the data accordingly. The project used a random 1% sample of the US census data from 2014. In order to compare the results, we evaluated both implementations (Kaggle's and non-experts' team) on the same data (Figure 8). The team of non-experts managed successfully to finish the project and their results were similar to to those published on Kaggle, resulting in the Pearson correlation coefficient ϱ=0.8.

The differences in the results can be traced back to the nuance that two implementations perform the data subsetting in different way. Each occupation in the data set is identified by a code. The 11 categories used in the project are quite generic, so it is user's responsibility to find the occupations which belong to the

[29] www.kaggle.com/wikunia/d/census/2013-americancommunity-survey/earnings-by-occupation-sex/

respective category. While Kaggle solution identifies only one occupation for each category, the Upwork team's implementation aggregates multiple occupation codes under the same category.

Task #2: Hillary Clinton's Emails[30]*:* This project explores the content of Hillary Clintons emails which were released by her in response to a Freedom of Information Act (FOIA) request and produces a heat-map of the countries that often appear in the emails. The dataset for this project is available on Kaggle. This project was also split into three assignments. The first assignment focused on data loading and cleaning, and consisted of three actions. The output of this task was a cleansed subset containing only the emails sent by Hillary Clinton and a list of all the countries in the world and their alternative spellings and abbreviations. The second task focused on identifying countries in the email data set and contained two actions – *Subset* and *Calculate occurrences*. The output of this task was a country occurrence list, containing the number of times each country is mentioned. The last task focused on the visualization part and consisted of two actions. The output was a sorted histogram and a heat-map in form of a world map, similar to the output of the Kaggle project.

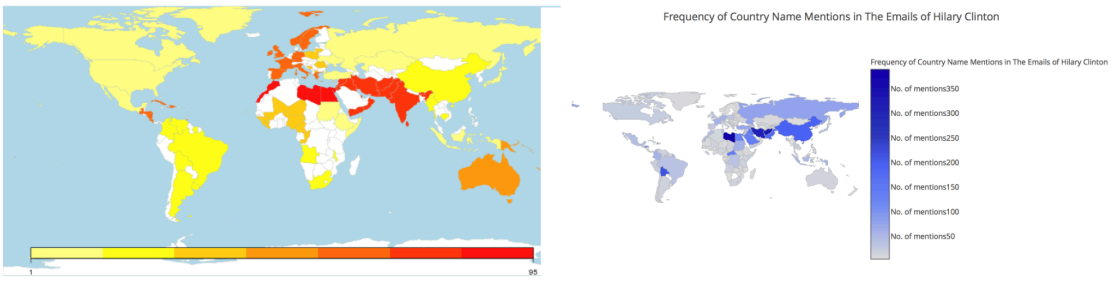

Figure 9.Pearson correlation coefficient ϱ= 0.72

The team of non-experts managed successfully to finish the project and the output of their work was similar to the results published on Kaggle (see Figure 9). In both implementations, the heat-map is based on a country occurrence list. We compared the results by calculating the Pearson correlation coefficient between the two lists with country occurrences which resulted in ϱ=0.72.

Similar to the previous project, the difference in the results is caused by the way two implementations identify the countries mentioned in the emails. The project on Kaggle and the team of non-experts use different approaches to identify countries abbreviations which leads to difference in the results.

Task #3: Reddit Sentiment Analysis[31]*:* The purpose of this project was to create a chart showing which Reddit comments receive the highest scores, based on the sentiment of the comment. Reddit is a large social network where users can submit content. The dynamics of this website is solely dependent on the number of up/down votes that the content receives. The content or comment with the highest number of votes is shown at the top. The categorization into three sentiment categories – objective,

[30] https://www.kaggle.com/ampaho/d/kaggle/hillary-clinton-emails/foreign-policy-map-through-hrc-s-emails/code

[31] https://www.kaggle.com/lplewa/d/reddit/reddit-comments-may-2015/communication-styles-vs-ranks/code

negative, and positive – was performed using the designated software package. The initial dataset includes Reddit comments from May 2015 and available on Kaggle.

The goal of *Reddit Sentiment Analysis* is to create a chart showing which Reddit comments receive the highest scores, based on the sentiment of the comment. Three sentiment categories were defined – objective, positive and negative. As in the previous project, we used a random sample of the May 2015 dataset. Both implementations were tested and evaluated using the same dataset. As it can be seen in Figure 10, the results are very similar – the average ranking scores for the positive, negative and objective comment categories are 6.18, 6,78, and 5.96 in the Kaggle project, and 5.75, 6.22, and 6.34 in the Upwork project performed by non-expert team.

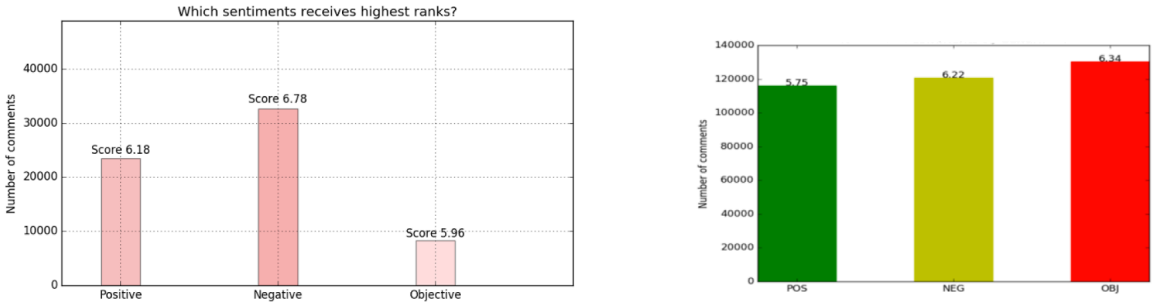

Figure 10 Equivalence tests: comparison of Kaggle with non-experts results

We also compared the ranking values in each sentiment category by performing equivalence tests on the results of the two projects (Mascha 2010). The goal of equivalence tests is to statistically test the equivalence of the variables. This was achieved by setting the equivalency region δ and testing whether the calculated confidence intervals for the differences between the two variables are within this region. For each sentiment category, we set the δ to be the average standard deviation of the Kaggle and the team of non-experts results. All the intervals are calculated with 95% confidence:

- Positive: CI_{poss} (-0.21, 1.07) ⊆ (-S.D.$_{poss}$, S.D.$_{poss}$)
- Objective: CI_{obj} (-1.16, 0.4) ⊆ (-S.D.$_{obj}$, S.D.$_{obj}$)
- Negative: CI_{neg} (-0.17, 1.29) ⊆ (-S.D.$_{neg}$, S.D.$_{neg}$)

In all cases the confidence intervals are contained within the equivalency region, meaning that there is no difference between the ranking means in each sentiment category.

Note that the implementation, the classification of the comments into one of the three sentiment categories, was done differently. In Kaggle project, the comments are classified by selecting only the comments with values above average (top quartile or top 3/8) for each sentiment, while the in project done by the non-expert team, sentiment scores are first normalized (through division by mean), and only then the comments are classified. Nevertheless, the results are almost identical.

Conclusions: At the end of this iteration, we qualitatively evaluated the features previously developed via a questionnaire, where we asked the participants open-end questions related to the use of the system. Specifically, we asked them to describe the features they found useful, difficulties they experienced in using the platform, and what are the functionalities that are missing or insufficient. We used

the feedback received in this iteration to improve our prototype and to add missing functionalities. For example, *we added a notification that the worker has finished his part such that the manager can review the output* and *the worker responsible for the next step can start working with the provisional results*. We also *added a notification to inform the owner of the notebook via email every time a "sticky-note" is attached*.

Regarding H2, all three experiments present substantial similarity between the experts' and non-experts' results. The similarity in the results of task #1 and task #2 is shown through significantly high correlation between the results – 0.8 and 0.72 correspondingly. Similarly, the results of task #3, compared using equivalence tests, indicate equivalence of the results. Altogether, the results of experiments support our hypothesis that crowds with mixed expertise are able to produce outputs comparable with the results produced by experts.

3.2.3 Third iteration – fully autonomous data analysis project

The last experiment we conducted was *Prediction in the Republican Primaries*[32]. The goal of this evaluation was to predict the results of the Republican Primaries 2015 in different counties. This experiment can be seen as full end-to-end data analysis project that includes all elements of data analysis, starting with data preprocessing, visualization, and up to building prediction models. The manager in this project, an expert worker from the crowd, was also responsible for building the prediction model. This setting allows the expert to better define the requirements of the activities, as he will use the processed data to build prediction models. In this project, the manager was responsible both for hiring the crowdworkers and defining assignments without intervention. Eventually, the project was split into three assignments performed by manager and two crowdworkers.
The first assignment focused on activities of *data loading, subsetting, and aggregating data* from different sources, such that the resulting data can be used for further analysis. The output of this task was a data-frame that included information about the primaries winner in every county and state as well as the demographic data of regions extracted from different data sources. This task required significant efforts and took about 5-7 hours of work. The second assignment was mainly about visualization of the data and descriptive statistics and resulted in various visualizations describing the relationship between population features of counties (e.g., residents' ethnicity or education, population density) and candidates' voting patterns. The duration of this task was about two hours. The last assignment was to build models predicting vote rates of each candidate. This task included training prediction models and testing them, similarly to Kaggle solution, on the test-set with reporting prediction qualities, such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The overall results of the prediction errors of the crowd and the experts are very similar. The mean absolute error of the Kaggle solution is $MAE_{Kaggle} = 6.5\%$, while the solution of the non-experts team yields $MAE_{Upwork}=6\%$. The root mean square errors of

[32] https://www.kaggle.com/apapiu/d/benhamner/2016-us-election/predictions-in-the-republican-primary

both solutions are almost identical with $RMSE_{Kaggle}$=8% in the Kaggle solution and $RMSE_{Upwork}$=7.7% in the model produced by crowdworkers.

Regarding H2 we can, hence, again conclude that the results produced by non-experts are comparable in their quality to those produced by data scientists.

4. Summary and discussion of results

Evaluation of H1: We tested the first hypothesis by reviewing the task decomposition output. Specifically, we aimed to ascertain whether it is possible to decompose the selected data analysis projects into sub-tasks such that the complexity of the sub-tasks is reduced compared to the overall complexity of the project. We asked the crowdworkers to report about the perceived complexity of the project and the sub-tasks. Following, we aggregated the results and analyzed them.

It was possible to split all projects into actions. Also, all of the workers were able to successfully complete their assignments. They rated the complexity of their assignment with an average of 2.25 (S.D.=0.96) out of 5. The project, on the other hand, was rated higher than the assignment complexity, with 2.42 out of 5 (S.D.=0.67). Despite the lack of significance (possibly due to the small sample size), we believe the results indicate a trend, that the method might work. Based on our evaluation and echoed by the literature review, we conclude that data analysis can be split into less complicated sub-tasks and accomplished by non-experts.

Evaluation of H2: To test the second hypothesis, we statistically compared the results of the projects conducted by experts with the results of non-experts that used our platform. As the data analysis projects we used for evaluation are publicly available on Kaggle, we explicitly asked the participants to not search and browse for the solutions on Kaggle. We also compared the code and the solutions' logic to assure that the code has not been inspired by the original solution. As already described in the iterations above, we attempted to cover a range of typical data analysis projects with complexity that meets real-world scenarios. Moreover, in order to ensure that the similarity is not a result of naturally limited space of solutions (which could lead to highly correlated results), we compared the results of other authors to see whether there is a natural variance in results.

Discussion: Both hypotheses have been empirically supported, meaning that data analysis projects can be effectively decomposed and accomplished with good quality. However, we found that the success of a project also greatly depends on other factors. The decomposed-tasks have to be effectively coordinated and timely adapted for the changing needs of data analysis. This is due to the dynamic/iterative nature of data analysis, where new insights, resulting from intermediate results, inspire new ideas on how to proceed with analysis. This, in turn, often requires additional data wrangling and sparks new iterations of work. While this work is performed in distributed way by non-experts, there is a need to support such process with appropriate coordination tool that will facilitate the process.

Moreover, the total cost of the experiments without hired manger was about 120 USD per project (the projects were split between three crowdworkers), where every worker has been paid 40 USD to accomplish her part, and each project required on average about 12 hours of work. In the project that involved the freelance manager, additional cost of 100 USD was paid for about 8 hours of manager's work. This

makes the projects economically competitive, especially in the light of the soaring data scientists' wage.

We also collected information about the background and skills of the crowdworkers that participated in our experiments. Most of them are bachelor or master students in their twenties, studying IT, computer or exact sciences, and working part-time as freelancers (13 hours per week on average). The workers perceive themselves mildly proficient in coding (self-rated with 3.2 out 5) and have basic background in data analysis, usually limited to introductory class in statistics or online course. Even though we have not conducted in-depth study on the demographics of online freelancers working in data analysis, our strong impression was that most of them can be characterized as part-time workers with average coding skills and very limited statistical/data analysis education with expected remuneration similar to the one in our experiments. This can be seen as evidence for the existence of sufficient talent to support the scenario we propose.

5. Limitations and future work

The proposed methodology has the following limitations. First the proposed top-down approach is not necessarily the optimal structure and other alternatives might be explored. For example, to allow workers to pick a task they want to work on in a self-managed manner and accompany the execution with managerial oversight. Second, we showed that the tasks can be decomposed into multiple simple sub-tasks. However, were not able to confirm this statistically. It is unclear whether this is due to a small sample of respondents (12). Future work might explore this by increasing the sample size and with recording additional data indicating the complexity of tasks. Third, to better evaluate the proposed platform, additional evaluation of the proposed scenario with other systems can be performed. For instance, the experiment where the coordination is done through a version control system that is used for software development such as GitHub[33]. Lastly, further research is needed to better understand the trade-off between the managerial overhead and saved costs due to outsourcing to non-experts.

6. Conclusion

This paper presents an approach of collaborative data analysis that involves data analysis novices with initial coding skills to participate in the process. We propose and evaluate the scenario where teams of non-experts are guided by expert throughout the process of data exploration and preprocessing. The proposed framework was evaluated with an especially designed tool and by virtue of multiple experiments, where the constraints are gradually released: first a pilot study, where we control for both the workers and the manager, then three experiments, where only the project manager is controlled, and ultimately, a data analysis project, where both the project manager and the workers are hired and perform the task without any external interference. The results demonstrate the feasibility of the proposed approach and support the hypothesis that the output of teams with mixed-level

[33] https://github.com/

expertise is equivalent to the results achieved by experts. Moreover, through various data analysis projects we show that it is possible to decompose them into simpler sub-tasks that can be then successfully accomplished by non-experts. Additionally, we found that the following features were valuable for collaborative data analysis with crowd workers: support for dynamic development, code deliberation, communication, and a journal with decisions made throughout the project.

In summary, we believe that our study paves the way for including non-expert crowd workers in data analysis tasks. As such, we hope to contribute to the research studying the requirements for building tools that can leverage the crowd to address the shortage of data analysts.

7. References

Affairs, L.W. and the T.F. on S.I. a P. a B. of S., 1999. Statistical methods in psychology journals. *American psychologist*, 54 (8)(8), pp.594–604.

Agrawal, A. et al., 2013. Digitization and the Contract Labor Market: A Research Agenda. *NBER Working Paper*, p.37.

Alasuutari, P., 2010. The rise and relevance of qualitative research. *International Journal of Social Research Methodology*, 13(2), pp.139–155.

Anderson, C., 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 16(7), pp.1–2. Available at: http://www.wired.com/print/science/discoveries/magazine/16-07/pb_theory.

Baker, M., 2016a. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), pp.452–454. Available at: http://www.nature.com/doifinder/10.1038/533452a.

Baker, M., 2016b. Statisticians issue warning on Pvalues. *Nature*, 531, p.151.

BARNES, W.H.F., 1944. The Nature of Explanation. *Nature*, 153(3890), pp.605–605. Available at: http://www.nature.com/articles/153605a0.

Bernstein, A., 2000. How can cooperative work tools support dynamic group process? Bridging the specificity frontier. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. pp. 279–288.

Bernstein, A., Klein, M. & Malone, T.W., 2012. Programming the global brain. *Communications of the ACM*, 55(5), p.41.

Bernstein, M.S. et al., 2010. Soylent: a word processor with a crowd inside. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp.313–322.

de Boer, P.M. & Bernstein, A., 2015. PPLib: Towards the Automated Generation of Crowd Computing Programs using Process Recombination and Auto-Experimentation. *ACM Transactions on Intelligent Systems and Technology*, (Special Issue: Crowd Computing).

Bollier, D., 2010. *The Promise and Peril of Big Data*, Available at: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.

Boyd, D. & Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, 15(5), pp.662–679.

Campbell, A. & Wu, A.S., 2011. Multi-agent role allocation: Issues, approaches, and multiple perspectives. *Autonomous Agents and Multi-Agent Systems*, 22(2), pp.317–355.

Campbell, J.L. et al., 2013. Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement. *Sociological Methods and*

Research, 42(3), pp.294–320.
Carpenter, J., 2011. May the best analyst win. *Science (New York, N.Y.)*, 331(6018), pp.698–699.
Chi, M.T.H., 2008. Three Types of Conceptual Change: Belief Revision, Mental Model Transformation, and Categorical Shift. In *Handbook of research on conceptual change*. pp. 61–82.
Collaboration, O.S., 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251), p.aac4716-aac4716. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716.
Conklin, E.J. & Yakemovic, K.C.B., 1991. A Process-Oriented Approach to Design Rationale. *Human-Computer Interaction*, 6, pp.357–391.
Creswell, J., 2002. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*,
Davenport, T.H. & Patil, D.J., 2012. Data_Scientist-the_Sexiest_Job_of_the_21St_Century.Pdf. , pp.70–76.
Davey Smith, G. & Ebrahim, S., 2002. Data dredging, bias, or confounding. *Bmj*, 325(7378), pp.1437–1438. Available at: http://www.bmj.com/cgi/doi/10.1136/bmj.325.7378.1437.
Van Dijck, J. & Nieborg, D., 2009. Wikinomics and its discontents: a critical analysis of Web 2.0 business manifestos. *New Media & Society*, 11(5), pp.855–874. Available at: http://journals.sagepub.com/doi/10.1177/1461444809105356.
Dissanayake, I., Zhang, J. & Gu, B., 2014. Virtual Team Performance in Crowdsourcing Contests : A Social Network Perspective. *ICIS 2015 Proceedings*, (Savage 2012), pp.1–16.
Dwork, C. et al., 2015. validity in adaptive data analysis. *Science*, 349(6248), pp.636–638. Available at: http://www.sciencemag.org/content/349/6248/636.
Edgell, S.E. & Noon, S.M., 1984. Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin*, 95(3), pp.576–583.
Eicken, H., 2013. Six red flags for suspect work. *Nature*, 497, pp.433--434.
Erceg-Hurn, D.M. & Mirosevich, V.M., 2008. Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), pp.591–601. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.63.7.591.
Fahy, P., 2001. Addressing some common problems in transcript analysis. *The International Review of Research in Open and Distance Learning*, 1(2), pp.1–6. Available at: http://www.irrodl.org/index.php/irrodl/article/viewArticle/321.
Feldman, M., Juldaschewa, F. & Bernstein, A., 2017. Data Analytics on Online Labor Markets: Opportunities and Challenges. Available at: http://arxiv.org/abs/1707.01790 [Accessed August 12, 2017].
Feldman, Mi., Anastasiu, C. & Bernstein, M., 2016. Towards Enabling Crowdsourced Collaborative Data Analysis. *Collective Intelligence*, (June), pp.1–5.
Fernandes-Taylor, S. et al., 2011. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC research notes*, 4(1), p.304. Available at: http://www.biomedcentral.com/1756-0500/4/304 [Accessed July 29, 2015].
Field, A., 2013. Discovering Statistics using IBM SPSS Statistics. *Discovering Statistics using IBM SPSS Statistics*, pp.297–321.
Fiske, S.T., 2016. How to publish rigorous experiments in the 21st century. *Journal of Experimental Social Psychology*, 66, pp.4–6. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022103116000032.

Fox, P. & Hendler, J., 2011. Changing the equation on scientific data visualization. *Science*, 331(6018), pp.705–708.

Friedkin, N.E. et al., 2016. Network science on belief system dynamics under logic constraints. *Science*, 354(6310), pp.321–326.

Gelman, A. & Hennig, C., 2015. Beyond subjective and objective in statistics. *arXiv preprint arXiv:1508.05453*. Available at: http://arxiv.org/abs/1508.05453 [Accessed September 16, 2016].

Gelman, A. & Loken, E., 2014a. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Psychological bulletin*, 140(5), pp.1272–1280. Available at: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf%5Cnhttp://doi.apa.org/getdoi.cfm?doi=10.1037/a0037714.

Gelman, A. & Loken, E., 2014b. The statistical Crisis in science. *American Scientist*, 102(6), pp.460–465.

Gelman, A. & Shalizi, C.R., 2015. Philosophy and the practice of Bayesian statistics Andrew. *British Journal of Mathematical and Statistical Psychology*, 66(1), pp.8–38.

Gilad-Bachrach, R., Navot, A. & Tishby, N., 2004. Margin based feature selection - theory and algorithms. In *Proceedings of the 21st International Confer- ence on Machine Learning*. p. 43. Available at: http://eprints.pascal-network.org/archive/00000869/.

Glaser, B.G. & Strauss, A.L., 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Available at: http://www.amazon.com/dp/0202302601.

Glass, G. V, Peckham, P.D. & Sanders, J.R., 2012. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance Author (s): Gene V . Glass , Percy D . Peckham and James R . Sanders Reviewed work (s): Source : Review of Educational Research , Vol . 42 , N. *Review of Educational Research*, 42(3), pp.237–288.

Good, P.I. & Hardin, J.W., 2012. *Common errors in statistics (and how to avoid them)*, John Wiley & Sons.

Gregor, S., 2006. The nature of theory in information systems. *MIS Quartely*, 30(3), pp.611–642.

Grolemund, G. & Wickham, H., 2014. A Cognitive Interpretation of Data Analysis. *International Journal of Statistics*, 82(2), pp.184–204. Available at: http://vita.had.co.nz/papers/sensemaking.pdf%5Cnhttp://onlinelibrary.wiley.com/doi/10.1111/insr.12028/abstract.

Gruber, T.R. & Russell, D.M., 1993. Generative Design Rationale: Beyond the Record and Replay Paradigm. *Design rationale: Concepts*, (December 1993). Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.1981&rep=rep1&type=pdf%5Cnpapers2://publication/uuid/AEA45CFD-89DF-4DE4-BC2D-71283E2E5DFB.

Guindon, R., 1990. Knowledge exploited by experts during software system design. *International Journal of Man-Machine Studies*, 33(3), pp.279–304.

Gutierrez, D.D., 2015. *Machine learning and data science: an introduction to statistical learning methods with R*, echnics Publications.

Haas, D. et al., 2015. Wisteria : Nurturing Scalable Data Cleaning Infrastructure. *Proceedings of the 41st International Conference on Very Large Data Bases*, 8(12), pp.2004–2007.

Head, M.L. et al., 2015. The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, 13(3).

Heer, J., Viégas, F.B. & Wattenberg, M., 2009. Voyagers and Voyeurs: Supporting

Asynchronous Collaborative Visualization. *Communications of the ACM*, 52(1), pp.87–97.
Hevner, A.R. et al., 2004. Design Science in Information Systems Research. *MIS quarterly*, 28(1), pp.75–105.
Hill, R.C. & Levenhagen, M., 1995. Metaphors and Mental Models: Sensemaking and Sensegiving in Innovative and Entrepreneurial Activities. *Journal of Management*, 21(6), pp.1057–1074.
Hoekstra, R., Kiers, H.A.L. & Johnson, A., 2012. Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3(MAY), pp.1–9.
Howison, J. & Crowston, K., 2013. Olla boration through open superposition.
Hruschka, D.J. et al., 2004. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods*, 16(3), pp.307–331. Available at: http://journals.sagepub.com/doi/10.1177/1525822X04266540.
Humphreys, M., Sanchez de la sierra, R. & Van der windt, P., 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1), pp.1–20.
Intelligence, A., 1992. Task-Structure Analysis Knowledge MOdeling for. , 35(9).
Introne, J. et al., 2013. Solving wicked social problems with socio-computational systems. *Kuntsliche Intelligenz*, 27(1), pp.45–52. Available at: http://cci.mit.edu/working_papers_2012_2013/cciwp2012-05colabkunstinel.pdf.
Ioannidis, J.P.A., 2005. Why most published research findings are false. *PLoS Medicine*, 2(8), pp.0696–0701.
Johnson-Laird, P.N., 1980. Mental models in cognitive science. *Cognitive Science*, 4(1), pp.71–115.
Jussim, L. et al., 2015. Interpretations and methods: Towards a more effectively self-correcting social psychology ☆. *Journal of Experimental Social Psychology*, xxx, pp.116–133. Available at: http://dx.doi.org/10.1016/j.jesp.2015.10.003.
Kalleberg, A.L. & Dunn, M., 2016. Good Jobs, Bad Jobs in the Gig Economy. *The Gig Economy: Employment Implications: Perspectives on Work 2016*, 20, pp.10–14.
Kandel, S. et al., 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. *Human factors in computing systems. ACM*, pp.3363–3372.
Kanji, G. k, 2006. *100 Statistical Tests* 3rd ed., London: SAGE Publications India Pvt Ltd. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006199-199501000-00015.
Kaptein, M. & Robertson, J., 2012. Rethinking Statistical Analysis Methods for CHI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.1105–1114. Available at: http://doi.acm.org/10.1145/2207676.2208557.
Kay, M., Nelson, G.L. & Hekler, E.B., 2016. Researcher-Centered Design of Statistics. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (August), pp.4521–4532. Available at: http://dl.acm.org/citation.cfm?doid=2858036.2858465.
Kittur, A. et al., 2011. CrowdForge: Crowdsourcing Complex Work. *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, pp.43–52. Available at: http://dl.acm.org/citation.cfm?doid=2047196.2047202.
Kittur, A. et al., 2012. CrowdWeaver : Visually Managing Complex Crowd Work. *Scenario*, pp.1033–1036. Available at: http://www.cs.cmu.edu/~pandre/pubs/crowdweaver-cscw2012.pdf.
Kittur, A., Nickerson, J. & Bernstein, M., 2013. The Future of Crowd Work. *Proc.*

CSCW '13, pp.1–17. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2190946%5Cnpapers2://publication/uuid/AE6BF263-1DEF-4900-8C95-DC8BAD2DE4AF.

Klein, G. & Moon, B., 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5), pp.88–92.

Klein, R.A. et al., 2014. Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), pp.142–152.

Krippendorff, K., 2004. *Content Analysis: An Introduction to Its Methodology*,

Krishnan, S. et al., 2015. SampleClean: Fast and Reliable Analytics on Dirty Data. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pp.59–75. Available at: http://sites.computer.org/debull/A15sept/p59.pdf.

Kulkarni, A., Can, M. & Hartmann, B., 2012. Collaboratively crowdsourcing workflows with turkomatic. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, p.1003. Available at: http://dl.acm.org/citation.cfm?doid=2145204.2145354.

Kurasaki, K.S., 2000. Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data. *Field Methods*, 12(3), pp.179–194.

Kuzon, W., Urbanchek, M.G. & Mccabe, S.J., 1997. Seven deadly sins of statistical analysis. *Journal of Oral and Maxillofacial Surgery*, 55(8), pp.897–898. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0278239197903773.

Lang, T. a & Altman, D.G., 2013. Basic Statistical Reporting for Articles Published in Biomedical Journals : The " Statistical Analyses and Methods in the Published Literature " or The SAMPL Guidelines ". *Science editors' handbook*, pp.29–32. Available at: http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf.

Langlois, R.N., 2002. Modularity in technology and organization. *Journal of Economic Behavior and Organization*, 49(1), pp.19–37.

Lee, J. & Lai, K.Y., 1991. What's in Design Rationale? *Humanï¿½Computer Interaction*, 6(3–4), pp.251–280.

Leek, J.T. & Peng, R.D., 2015. P values are just the tip of the iceberg. *Nature*, 520(7549), p.612.

Lukacs, P.M., Burnham, K.P. & Anderson, D.R., 2010. Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, 62(1), pp.117–125.

MacDonald, J., 2003. Assessing online collaborative learning: Process and product. *Computers and Education*, 40(4), pp.377–391.

MacLean, A. et al., 1991. Questions, Options, and Criteria: Elements of Design Space Analysis. *Humanï¿½Computer Interaction*, 6(3–4), pp.201–250.

Malone, T.W. et al., 1999. Tools for Inventing Organizations : Toward a Handbook of Organizational Processes Tools for Inventing Organizations : Toward a Handbook of Organizational Processes. , 3(May 2015), pp.425–443.

Mann, M., 2016. Must try harder. *New Scientist*. Available at: http://www.sciencedirect.com/science/article/pii/S0262407916303682 [Accessed August 30, 2016].

Martin Bland, J. & Altman, D., 1986. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet*, 327(8476), pp.307–310.

Mascha, E.J., 2010. Equivalence and noninferiority testing in anesthesiology research. *Anesthesiology*, 113(4), pp.779–781.

Miles, M., Huberman, M. & Saldana, J., 2014. *Qualitative Data Analysis*,

Morton, K. et al., 2014. Support the Data Enthusiast: Challenges for Next-Generation

Data-Analysis Systems. *Proceedings of the VLDB Endowment, Volume 7, pp. 453–456, 2014*, 7, pp.453–456. Available at: http://homes.cs.washington.edu/~kmorton/p446-morton.pdf.

Nimon, K.F., 2012. Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, 3(AUG), pp.1–5.

Van Noorden, R., 2014. Online collaboration: Scientists and the social network. *Nature*, 512(7513), pp.126–129. Available at: http://www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711.

Norman, D.A., 1983. Some Observations on Mental Models. In *Mental Models*. pp. 7–14. Available at: http://www.amazon.com/Mental-Models-Cognitive-Science-Series/dp/0898592429.

Nosek, B.A., Spies, J.R. & Motyl, M., 2012. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, 7(6), pp.615–631. Available at: http://pps.sagepub.com/lookup/doi/10.1177/1745691612459058.

Nuzzo, R., 2014. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487), pp.150–152.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251), p.aac4716-aac4716. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26315443.

Osborne, J. & Waters, E., 2002. Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, 8(2), p.1.

Ott, E.M., 1989. Effects of the Male-Female Ratio at Work: Policewomen and Male Nurses. *Psychology of Women Quarterly*, 13(1), pp.41–57.

Paglieri, F., 2004. Data-oriented belief revision: Towards a unified theory of epistemic processing. *Proceedings of STAIRS*. Available at: http://books.google.com/books?hl=en&lr=&id=Z569jqwQuK8C&oi=fnd&pg=PA179&dq=Data-oriented+Belief+Revision+:+Towards+a+Unified+Theory+of+Epistemic+Processing&ots=SqAEHHjdec&sig=Out0eaWHx3vygt_eBCzHm6QJYcA.

Partington, D., 2013. *Essential Skills for Management Research*,

Peffers, K. et al., 2008. A Design Science Research Methodology for Information Systems Research. *J. Manage. Inf. Syst.*, 24(January), pp.45–77.

Ransbotham, S., Kiron, D. & Prentice, P.K., 2015. The Talent Dividend. *MIT Sloan Management Review*, 56(4), pp.1–12. Available at: http://sloanreview.mit.edu/projects/analytics-talent-dividend/.

Redmiles, D., 2000. Software Requirements for Supporting Collaboration through Categories.

Reinecke, K. & Bernstein, A., 2013. Knowing What a User Likes: A Sesign Science Approach to Interfaces that Automatically Adapt to Culture. , 37(2), pp.427–453.

Rouder, J.N. et al., 2016. Is There A Free Lunch In Inference? *topiCS*, 8(1), pp.1–5.

Russell, D.M. et al., 1993. The cost structure of sensemaking. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93*. pp. 269–276. Available at: http://portal.acm.org/citation.cfm?doid=169059.169209.

Russo, D. & Zou, J., 2016. Controlling Bias in Adaptive Data Analysis Using Information Theory. In *Proceedings of the 19th International Con- ference on Artificial Intelligence and Statistics (AISTATS) 2016*. pp. 1232–1240. Available at: http://arxiv.org/abs/1511.05219.

Saldana, J., 2011. *Fundamentals of Qualitative Research: Understanding Qualitative Research*,

Salehi, N. et al., 2016. Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

dos Santos, F. & Bazzan, A.L.C., 2009. An ant based algorithm for task allocation in large-scale and dynamic multiagent scenarios. *Proceedings of the 11th Annual conference on Genetic and evolutionary computation - GECCO '09*, p.73. Available at: http://portal.acm.org/citation.cfm?doid=1569901.1569912.

Van Schaik, P. & Weston, M., 2016. Magnitude-based inference and its application in user research. *International Journal of Human Computer Studies*, 88(August), pp.38–50.

Schlauderer, S. & Overhage, S., 2013. Exploring the Customer Perspective of Agile Development: Acceptance Factors and on-Site Customer Perceptions in Scrum Projects. *Thirty Fourth International Conference on Information Systems*, pp.1–20.

Schubanz, M., 2014. Design rationale capture in software architecture: What has to be captured? In *WCOP 2014 - Proceedings of the 19th International Doctoral Symposium on Components and Architecture (Part of CompArch 2014)*. pp. 31–36. Available at: http://dx.doi.org/10.1145/2601328.2601329.

Sculley, D. & Pasanek, B.M., 2008. Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, 23(4), pp.409–424.

Seel, N.M., 2001. Epistemology, situated cognition, and mental models: "Like a bridge over troubled water." *Instructional Science*, 29(4–5), pp.403–427.

Seitz, F., Heisenberg, W. & Pauli, W., 2000. Decline of the generalist The vigour of every discipline depends on people of broad vision . *Nature*, 403(February), pp.10021–10021.

Sere, F.C. et al., 2011. Online collaboration: Collaborative behavior patterns and factors affecting globally distributed team performance. *Computers in Human Behavior*, 27(1), pp.490–503.

Sheskin, D.J., 2004. Handbook of parametric and nonparametric statistical procedures. *Technometrics*, 46, p.1193. Available at: http://books.google.com/books?id=bmwhcJqq01cC&pgis=1.

Silberzahn, R. & Uhlmann, E.L., 2015. Many Hands Make Tight Work. *Nature*, 526(7572), pp.189–191. Available at: http://www.nature.com/news/crowdsourced-research-many-hands-make-tight-work-1.18508.

Simmons, J.P., Nelson, L.D. & Simonsohn, U., 2011. False-Positive Psychology. *Psychological Science*, 22(11), pp.1359–1366. Available at: http://journals.sagepub.com/doi/10.1177/0956797611417632.

Smith, A.J., 1990. The Task of the Refereee. , pp.1–7.

Stefik, M., 1981. Planning with constraints (MOLGEN: Part 1). *Artificial Intelligence*, 16(2), pp.111–139.

Stein, R.T. & Heller, T., 1979. An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *Journal of Personality and Social Psychology*, 37(11), pp.1993–2002.

Strasak, A.M. et al., 2007. Statistical errors in medical research - A review of common pitfalls. *Swiss Medical Weekly*, 137(3–4), pp.44–49.

Strauss, A. & Corbin, J., 1990. Basics of qualitative research: grounded theory procedure and techniques. *Qualitative Sociology*, 13(1), pp.3–21.

Thomas, D.R., 2006. A General Inductive Approach for Analyzing Qualitative

Evaluation Data. *American Journal of Evaluation*, 27(2), pp.237–246. Available at: http://journals.sagepub.com/doi/10.1177/1098214005283748.

Tseng, H. et al., 2009. Key Factors in Online Collaboration and Their Relationship to Teamwork Satisfaction. *The Quarterly Review of Distance Education*, 10(626), pp.195–206.

Tukey, J.W. & Wilk, M.B., 1966. Data analysis and statistics: an expository overview. *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference*, (695), pp.695–710. Available at: http://dl.acm.org/citation.cfm?id=1464366.

Vargha, A. & Delaney, H., 1998. The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, 23(2), pp.170–192.

Viegas, F.B. et al., 2007. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), pp.1121–1128.

Weiss, G. & Wodak, R., 2003. *Critical Discourse Analysis*,

Westfall, J. & Yarkoni, T., 2016. Statistically controlling for confounding constructs is harder than you think. *PloS one*, 11(3), pp.1–22.

Willett, W. et al., 2011. CommentSpace: Structured Support for Collaborative Visual Analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.3131–3140. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.188.1845&rep=rep1&type=pdf.

de Winter, J.C.F. & Dodou, D., 2010. Five-Point Likert Items : t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11), pp.1–16. Available at: http://pareonline.net/pdf/v15n11.pdf.

Woolston, C., 2015. Psychology journal bans P values. *Nature*, 519(7541), pp.9–9. Available at: http://www.nature.com/doifinder/10.1038/519009f [Accessed August 12, 2017].

Yadav, M.S. & Pavlou, P.A., 2014. Marketing in Computer-Mediated Environments: Research Synthesis and New Directions. *Journal of Marketing*, 78(1), pp.20–40. Available at: http://journals.ama.org/doi/abs/10.1509/jm.12.0020.

Yukl, G., 2001. Leadership in organizations. *Personnel Psychology*, 7th(4), p.542. Available at: http://files.liderancaecoaching.webnode.com/200000015-31f5732fb3/media-F7B-97-randd-leaders-business-yukl.pdf.

Zimmerman, D.W., 2004. Inflation of Type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation tests. *Psicologica*, 25(1), pp.103–133.

Zimmerman, D.W., 1998. Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. *The Journal of Experimental Education*, 67(1), pp.55–68.

Analysis of Behavioral Factors Underlying the Data Analysis Process

This chapter is based on a paper that is to be submitted to one of journals concerned with improving the general scientific process

Analysis of Behavioral Factors Underlying the Data Analysis Process

Abstract

When scientists analyze data they are confronted with a myriad of subjective decisions that may or may not impact the results. In this crowdsourced project, many analysts used the same complex dataset to test the same predefined hypotheses regarding the role of gender and status in academic debates. Using specially designed platform called Data Explained, we obtain fine-grained information reflecting the rationale for each step undertaken during the data analysis process. We apply a General Qualitative Approach to identify factors underlying the variability in data analysis choices and discuss how these degrees of freedom could be mitigated or made transparent. Our study contributes to better understanding the behavioral factors underlying the data analysis process and to the ongoing discussion of the reproducibility crisis in science.

1. Introduction

The recent crisis of confidence in science has called existing research approaches into question. Recent attempts to reproduce published findings across different fields led to the surprising conclusion that most of the studies published in top journals can not be easily reproduced (Collaboration 2015; Ioannidis 2005). There are various reasons for this, such as lack of access to the original data, incomplete reporting of assumptions made during the analysis (see Feldman et al. manuscript I), and subjective decisions which are made by researcher and not always reflected in the final report.

Some of the difficulties in reproducing scientific findings may not have anything to do with the experimental method per se, but with the way in which data is typically analyzed. Explicit or implicit assumptions about the investigated hypothesis might lead to variability in way the data is preprocessed and in the chosen analytic approach (Sculley & Pasanek 2008). In a typical scientific article, a single researcher or team of researchers present their analysis of a dataset they have collected. However, there are often numerous analytic strategies that could all be plausible alternatives when analyzing the same dataset, and variations in these strategies could produce very different outcomes, a process referred to as the so-called "garden of the forking paths" (Gelman & Loken 2014a). To explore the extent to which possible forking paths influence specific aspects of the data analysis, we provided a dataset to many analysts and asked them to test two hypotheses with that same dataset while carefully tracking every decision using an online platform we developed called Data Explained. By doing so, we are able to observe the roadmap of different analytical alternatives and decisions in much greater detail than ever before. This approach elicits the major factors that underlie the diversity in the analytical process: a situation where researchers reach different results when they analyze the same data and test the same hypothesis (Silberzahn et al., in press). Specifically, we analyze the steps undertaken by data analysts and explore factors underlying the implicit decisions made throughout their data analyses. When similar results are obtained by many analysts who are blind to each others' approaches, scientists can speak with one voice on an issue. When different analysis

strategies converge to very similar estimated effects, it indicates robust results despite variation in analysis strategies. In contrast, if the observed effects are highly contingent on subjective analytic decisions or varying analysis strategies, then the results are more equivocal. Hence, a crowdsourcing approach offers a unique opportunity to identify variability in data analysis strategies and the particular approach we took allowed us to observe sources of this variation in more detail than ever before.

To explore the latent factors underlying subjective decisions, we rely on a general qualitative approach to analyze the explanations provided by different analysts. Following this approach, a team of researchers analyzed the descriptive text explaining in detail every step undertaken by analysts throughout data analysis as well as the source-code corresponding to each step in data analysis. To examine the exact points at which the paths diverged and forked off, we developed a web platform called DataExplained that allows analysts to conduct their data analysis online whilst explaining at every step their decisions as they progress in the analysis. By asking analysts to explain their decisions and considered alternatives to the executed code, we obtain a rich dataset capturing the various workflows of the different analysts. This is especially useful due to the exploratory nature of data analysis, where analysts often experiment with data prior to deciding on how to proceed with analysis.

We asked the participants to independently analyze a dataset of intellectual conversations from Edge.org that was collected in 2015 and ranged from 1996 to year 2014. The dataset contained 123 edge conversations, with 60 attributes related to the conversation, its participants or the textual level of the transcript. Attributes not provided on the website were manually collected by browsing CVs, university or personal web-pages, and professional networking websites. A detailed procedure of every step followed during the creation of the dataset, along with a full description of every attribute, can be found in **Appendix A.1**.

The participants were asked to test the following hypotheses using the dataset:

- Hypothesis 1: "A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion."
- Hypothesis 2: "Higher status participants are more verbose than are lower status participants."

The operationalization of key variables (e.g., female participation, researcher status in the professional hierarchy, dominant language) was left unconstrained and up to the individual researchers. Every analyst could choose to use dataset variables such as citation counts (possibly corrected for self-citations), publications in high-impact journals, tenure status, ranking of current university, ranking of doctoral institution, years since PhD, or some combination of the above as indices of a researcher's' status (see appendix A.1). Analysts could also choose to focus on status within a field, subfield, or among participants in an individual conversation. Likewise, dominant language was likely to be operationalized using different word lists in automated text analyses by different researchers. Thus we sought to capture the ambiguity research analysts typically face when approaching a complicated dataset and choosing how to operationalize their variables of interest.

We analyzed the collected meta-data based on the way the variables were operationalized, what statistical methods were applied, and how particular variables were taken into account, all leading to the diversity in the results.

Given the above, the major contributions of this study are the following:

- An exploratory study of the major factors that underlie the variability in data analysis.
- A proposed model that conceptualizes cognitive processes involved in data analysis in systematic way.
- A web-based platform that supports transparent data analysis reporting. The platform records all executed source-code and prompts analysts to comment on their code and analytical thinking steps. The platform also allows to graphically represent the workflow of analysis.

The paper is structured as follows: the literature review is followed by sections describing the methodology and the research design. We then report the results of our study where we outline the major factors accounting for data variability and propose a model describing their interplay during data analysis. Lastly, we discuss the results in the light of the crisis of confidence in science and propose how crowdsourcing data analysis might make transparent how subjective choices affect research results.

2. Literature Review

In this section we first review the literature on variability in statistical results, followed by the studies addressing possible solutions**.** We then provide theoretical background to the cognitive theory of sensemaking and describe the literature discussing the cognitive perspective in data analysis. Finally, we review the field of design rationale and explain how it can be used as a potential tool to explain variability in data analysis.

2.1 Variability in Statistical Analyses

Null Hypothesis Significance Testing (NHST) is often used by researchers to test hypotheses. This is based on the assumption that if the difference between two means, which are assumed under the null-hypothesis and alternative hypothesis is significant according the to a threshold probability (i.e. significance level), the null hypothesis can be rejected. A p-value measure is used to quantify the probability for this deviation in order to decide whether to reject the null-hypothesis or not. Gelman & Loken (2014b)point out that the hypothesis can often be operationalized or tested in many ways, even such that statistically significant results emerge. Gelman and Loken call this issue the *multiple comparison problem* – also widely known as "p-hacking", "researcher degrees of freedom" or "selective reporting" (Head et al. 2015; Simmons et al. 2011) However, p-hacking, to certain extent, assuming intent in research conduct which is often not the case. On the contrary, researchers often find themselves contemplating on how to proceed as the problem they research is ambiguous by its nature and might be approached differently with different operationalization and assumptions in mind. In such case, the nature of the problem likely to lay not in intentionally biased approach of the

researcher, but in the inherent ambiguity of the phenomenon under investigation, personal characteristics of the researchers, and interactions between the two.

But even in cases where researchers follow accepted procedures of ex-ante fixing their hypotheses and variables, data analyses may till differ in the way data is processed and in preferred statistical methodology. Moreover, different workflows in data-analysis, possibly due to implicit decisions researchers make during their analysis, may lead to variance in the results. Gelman and Loken (2013) conventionally call this variability in potential ways of analyzing data while having a hypothesis in mind "a garden of forking paths." Moreover, authors note that although some paths may lead to statistically significant results, it is wrong to conclude that the presented evidence of the initially formed alternative hypothesis is true. For instance, consider the issue of method selection. The classic scientific approach assumes that the method for data analysis is selected independently of the data, and before the analyst has explored it. In practice, the method is typically selected as a function of (or after investigating the) data at hand. This approach is called *adaptive data analysis*. If a data analyst consciously prefers a certain model which is more likely to produce a desired outcome, the selected model could be described as a "fished model" (Humphreys et al. 2013). This behavior might be mostly attributed to implicit decisions and judgments from analysts (Dwork et al. 2015; Gelman & Hennig 2015). On the other hand, certain method might be selected not as a result of malicious intent but as a result of personal method preference (e.g. Bayesian vs. Frequentist method). Silberzahn and Uhlman (in press) introduce a term "analysis contingent results" to describe how defensible, but subjective decisions impact analysis results. The uncertainty regarding the best path to proceed with, as well as the researcher's background knowledge and preferences might be part of the basis of differences in results.

Another setting where the role of implicit decisions surfaces is the "curse of dimensionality" (aka Freedman's paradox) – a multidimensional dataset where the number of explanatory variables is very high compared to the number of entries in the dataset (Bellman, 2013; Freedman, 1983). In such data, some of the variables might be highly correlated by chance. As a consequence, this can lead to false confidence in the predictability of some explanatory variables. As Lukacs et al. (2010) point out, this paradox is an extreme case of model selection bias, as the effect of slightly correlated explanatory variables are overestimated. Possible measures to account for presumably high correlations are R, Mallows's Cp, R adjusted, AIC or BIC, which penalize number of explanatory variables. In this context another problem that might arise when there is a need to perform feature selection - deciding on the variables that will be included in the analysis - is biased selection of the variables for further analysis. While there are exist certain methods for evidence driven variables selection (e.g. Gilad-Bachrach et al. 2004) it is not uncommon for researchers to select variables that seem to be most informative and interesting for the analysis.

Frequently, datasets are reused multiple times, while the results and insights derived from previous studies often inform subsequent analyses. Arguably, this may bias outcomes since signals revealed in previous studies may bias the the forthcoming study (Russo & Zou 2016; Dwork et al. 2015). Moreover, given the rise of the Big Data phenomenon, data analysts are confronted with increasingly

complex data consisting non-trivial relationships, which leaves more room for subjective decisions. For instance, possible relationships among variables may not be easily visible anymore if the amount of data is too big to make sense of it at first glance. This can lead to the practice of apophenia: seeing patterns in the data, although they do not actually exist. therefore, it is easy to fall into a trap of identifying pseudo-signals by observing noisy (big) data that occurs just by chance (Boyd & Crawford 2012). Further, as Bollier (2010) points out, cleaning a large amount of raw data often presents problems in maintaining an objective interpretation of the data – especially if data originates from disparate sources. As a consequence, subjective assumptions have to be made to link multiple datasets together. To mitigate this challenge, it is important to build a model which represents the data originating from different sources in its respective context *prior* to integration, to avoid misinterpreting correlation as causation(Anderson 2008). Regardless with the size of the dataset however, an analysis is always subject to limitations and bias(Boyd & Crawford 2012).

2.2 Data analysis: a cognitive perspective

As researchers conduct data analyses, they obtain intermediate results. These results are almost always interpretative in their nature and often stem from personal understanding and beliefs, which may vary across individuals. Since data analysis is an iterative process, intermediate output plays a key role in deciding which path to further follow. Thereby, a data analysis not only incorporates statistical or computational steps, but also cognitive processes. As Grolemund and Wickham, (2014) point out, "*data analyses rely on the mind's ability to learn, analyze, and understand*", where each data-driven scientific work aims *to "educate a reader about some aspect of reality"*. These readers may have different professional backgrounds and/or experiences in data analysis, as well as different mental frameworks for dealing with such tasks (i.e. forming mental models).

The concept of mental models has been studied in various research areas of cognitive science for many years(e.g. Barness 1944; Johnson-Laird 1980; Norman 1983; Seel 2001; Weiss & Wodak 2003). Scientists describe it as "*subjective representation of the events, action, or situation a discourse is about*" (Weiss & Wodak 2003) or "*qualitative mental representations which are developed by subjects on the basis of their available world knowledge aiming at solving problems or acquiring competence in a specific domain*"(Seel 2001). The process of building and interpreting such descriptions of mental models or schemes is also known as *sensemaking* (Russell et al. 1993). Being confronted with data, situated cognition and reasoning in the sensemaking process have a considerable influence on how the data is interpreted and transformed into summary results and conclusions. Prior beliefs about a certain phenomenon may be absent, incomplete, or conflict with the apparent empirical results. Information gained from the data can help fill such gaps (if prior beliefs are incomplete), expanded (if prior beliefs are missing) or even revised (if false prior beliefs are contradicting correct information),(Chi 2008). Hence, the data by itself can influence an analyst's beliefs, which, as a consequence, leads to different analytical choices(Paglieri 2004). A possible tool that can help researchers explore complex data and build better intuitions are appropriate visualizations(Morton et al. 2014; Fox & Hendler 2011). Without the need of knowledge for specific programming or

query languages, visual analytics might serve as efficient sensemaking tool. When being confronted with a lot of data, visualizations or visual exploration tools might help to make sense of the interplay between multiple datasets. Especially when the data is of dynamic nature (e.g. temperature profiles), appropriate visualizations can help data analysts reveal new substantial patterns, which in turn might lead to adaptations of beliefs and/or mental models (Bollier 2010).

That cognitive processes play a key role in data analysis has been acknowledged by some leading statisticians. Tukey and Wilk (1966) describe exploratory data analysis as the "intent to seek through a body of data for *interesting relationships and information* and to exhibit the results in such a way as to make them *recognizable to the data analyzer*". What would be the interesting information and relationships in such case? What information is recognizable by data analyst and what will be overlooked? – this is likely to be contingent on data analyst's perception, agenda, as well as various extrinsic constraints. Moreover, authors state that at all stages of the data analysis process the outcomes of data analysis, would it be actual or potential results, have to be matched to the capabilities of people analyzing it. This way, successful data analysis is subject to the ability to process and understand the results. Moreover, even "black box" data analysis methods like deep learning, which has gained a recognition in the recent years, is not useful unless the analyst can meaningfully interpret the results. Such ability relies not within the professional or technical ability but is part of a cognitive process inherent to the research process(Grolemund & Wickham 2014).

2.3 Design rationale: Capturing the factors underlying analysis contingent results

Once decided which course to take throughout data analysis, it is of interest to explain the rationale behind this decision. Why should one follow this exact path, or why is this the right path to follow? Hill and Levenhagen (1995) describe this (implicit) action of communicating the perceived mental model as *sensegiving*, which eventually results in shared belief systems or consensuses(Friedkin et al. 2016). The description of the motivations underlying decisions in the context of designing a system or artefact, is also referred to as Design Rationale (DR)(Lee & Lai 1991). DR can be defined as "*[...] explanation of why an artifact is designed the way it is*". Along with many other research areas, DR is widely discussed in the field of computer science(Schubanz 2014; Gruber & Russell 1993). Especially in software development, it can help to effectively document and maintain artefacts (from both, the UI designer's point of view, as well as the technical engineer's perspective)(Guindon 1990). The classic concepts of a design rationale system include the existence of a design rationale database (containing design histories, reasoning, decisions, etc.). This database can be accessed with an appropriate representation schema, which elicits argumentations, decisions, or advantages and disadvantages of different options. In our case, an analyst implicitly accesses this system during the sensemaking/sensegiving processes. This reflects the definition of Conklin and Yakemovic (1991), which say that DR can be seen as the path of decisions and selected alternatives that join the initial state (in which no decisions have been made) to the final state (in which all design decisions have been resolved). Following the metaphor of a garden with forking paths, one could argue that any data exploration is like walking within the garden with tangled paths that might lead to different

exits. In this metaphor, one could say that DR represents the full explanation as for why a certain path was preferred over others. We can describe each sub-path as a cognitive cycle a data analyst traverses, since at every of these forks the analysts repeatedly revisit and revise their beliefs and mental models.

3. Methodology

In this section we first describe the platform we designed to track scientists' data analytic decisions as they made them. We then describe the overall methodology of the crowdsourced initiative and outline our analysis process for the meta-scientific data generated by the project.

3.1 Analysis Platform: DataExplained

To conduct the experiment, we designed an online platform, DataExplained, that allows participants to run an analysis online in a RStudio environment. The platform's core consists of RStudio Server, which allows participants to conduct a data analysis using RStudio via web browser. In addition to the online RStudio environment, we implemented features that enabled us to track all executed commands along with the analysts' detailed explanations for every step of the executed analysis. The procedure used was as follows:

1. The participants were provided access to the platform, where they executed their data analysis using the RStudio user web-interface. During their analysis, every executed command (i.e. log) was recorded. Recording all executed commands (i.e. commands executed but not necessarily found in the final code) is useful, as such logs might reveal information that affected the analysts' decisions but are not reflected in the final script. Whenever the participants believed that a series of logs can be described as a self-explanatory block, or when a certain number of logs was produced, they were asked to describe their rationales and thoughts about the underlying code.

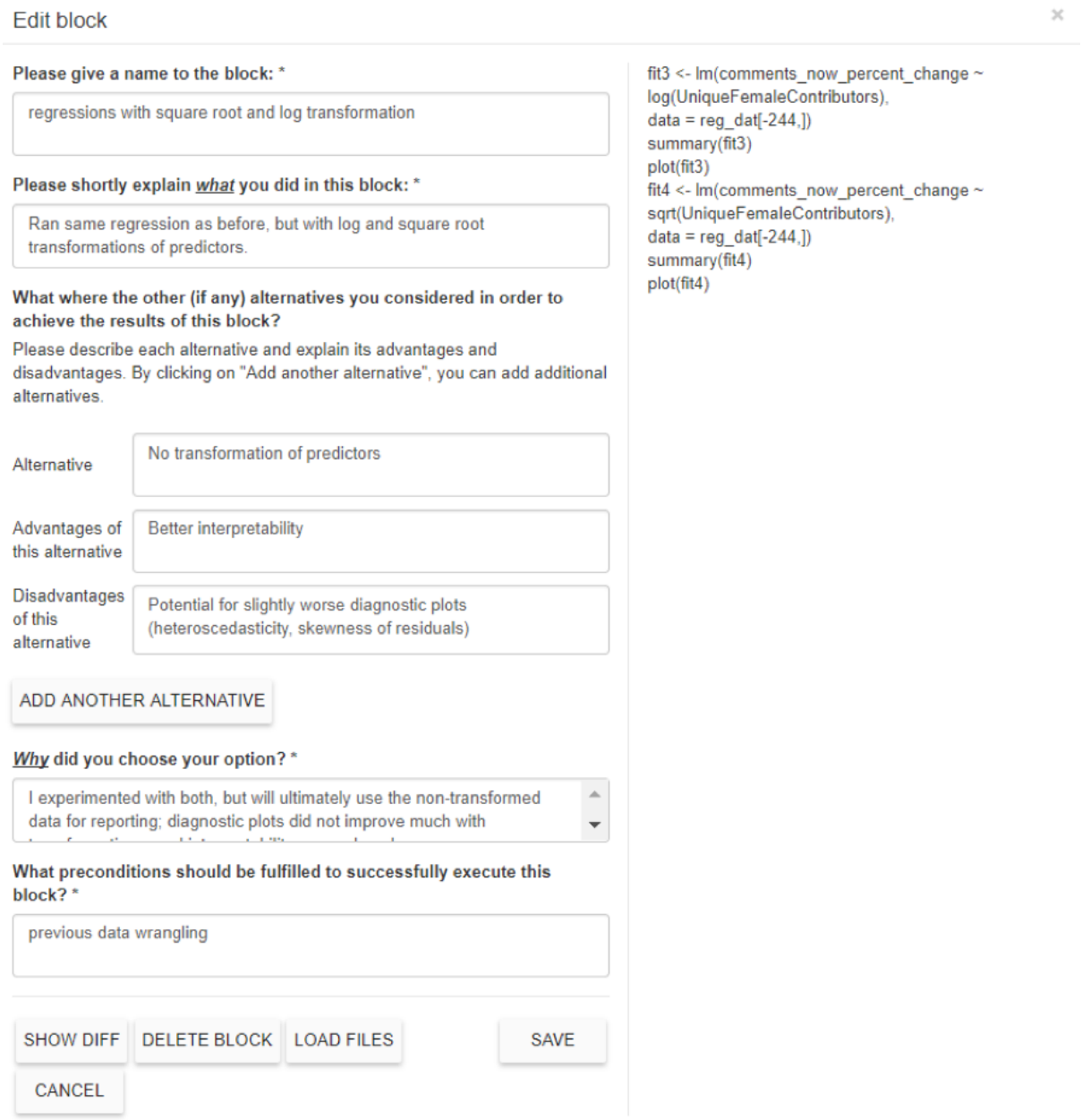

Figure 11. A block of logs with the explanations for the code.

Each block (see Figure 11) consisted of a few questions:

- Please shortly explain what you did in this block?
- What preconditions should be fulfilled to successfully execute this block?
- What were the other (if any) alternatives you considered in order to achieve the results of this block?
    - Explain the alternative
    - Explain the advantages
    - Explain the disadvantage
- Why did you choose your option?

By answering these questions, we were able to allow the analysts to communicate the reasoning underlying their data analysis. This allowed us to observe the reasons

underlying an analytic decision, the justification for it, the considered alternatives, the trade-offs evaluated, and the deliberation that led to the final implementation.

To help participants recall any recent changes in code, we embedded a system where it is possible to visually explore the code differences between the subsequent blocks. Additionally, participants were able to navigate through their analysis history, by restoring the state of the RStudio workspace at any given point a block was created. These features helped the analysts to recall the considerations during their analysis, even if the corresponding part of code did not exist anymore in the final script.

Second, the analysts were provided with an overview of all blocks that they created during their data analysis. They could edit the blocks and reassign the respective logs to other blocks. This might be desirable, if a block is not reflecting the originally anticipated goal anymore. It also allowed them to read the description of blocks following a storyline and edit the current description accordingly. At this stage, it is also possible to create new blocks that will better reflect an analyst's line of thought.

Finally, in the last step of data analysis using DataExplained analysts are asked graphically to model the workflow representing the evolution of the analysis. Initially, each analyst is presented with a straight chain of blocks, ordered by their execution. The analysts are then asked to restructure the workflow such that it better reflects the actual process. For example, iterative cycles of trying out different approaches for a sub-problem could be modeled as loops in the workflow (c.f. example of workflow visualization from one of the participants in Figure 12).

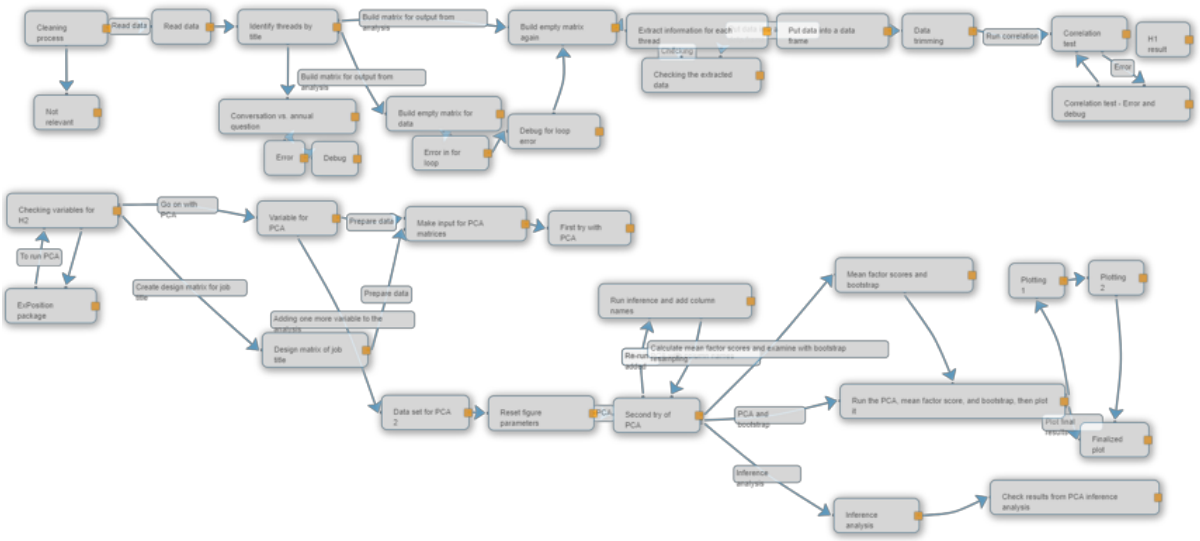

Figure 12. Snippet of workflow modelled by a participant

Upon completion of the experiment, analysts responded to a survey in which they were asked to report their empirical results (i.e. their estimated effect sizes for the two focal hypotheses), the applied methods, and a short assessment of their (possibly updated) beliefs regarding the two hypotheses.

In this study we follow a qualitative research approach, which is suited for the research that relies on non-structured data to describe social phenomena(Alasuutari 2010). As described by Thomas (2006), there are four major approaches for qualitative analyses: Discourse Analysis, Grounded Theory, Phenomenology, and

the General Inductive Approach. Subsequently, we briefly present these approaches and explain our methodological choice.

In the social sciences, *discourse analysis* usually focuses on analyzing text as a mean of eliciting social practices and rhetoric which are emerging around topics of interest. *Phenomenology* seeks to understand the personal experiences of people who share the same experiences. The result is a coherent story describing the studied phenomenon based on the multifold of individual perspectives. The goal of a *Grounded Theory* approach is to generate a theory using a bottom-up approach based on axial coding and theoretical sampling. Last but not least, a *General Inductive Approach* seeks to develop a framework of the underlying structure of experiences or processes that are evident in the raw data. The primary goal is to allow research findings to emerge from the frequent, dominant, or significant themes inherent in raw data, without the constraints imposed by structured methodologies. This approach is more lightweight and it can lead to reliable and valid findings by following a set of standardized procedures. Even though this method is not as well-rooted as other approaches for theory building (such as Grounded Theory), it is well accepted as feasible approach to answer research questions about understanding the underlying process.

In this study, we follow the General Inductive Approach mainly for the following reasons: the classical Grounded Theory approaches coined by Glaser et al. (1967) as well as Strauss and Corbin (1990) are restrictive in terms of rules and procedures to follow, and often not straightforward (Partington 2013; Thomas 2006). This approach limits the inductive learning process to be entirely isolated from any impact of existing theories. However, since the cognitive aspect of data analysis has been recognized in literature for long time and because the phenomenon of variability in data analytic approaches has been described in the recent literature, we intend to draw from the existing literature. We therefore adopted a less restrictive framework for our study. The General inductive approach is the most suitable for this meta-scientific project, as it allows us to follow the bottom-up approach of inferring key factors and at the same time allows to draw on existing theories such as sensemaking.

3.2 The General Inductive Approach

Inductive (qualitative) coding is central to the General Inductive Approach and usually applied when there is a need to analyze volumes of verbal and written material in order to identify patterns and gain insights about the research problem. The process starts with (usually) multiple researchers carefully reading the relevant materials and considering possible meanings reflected in the text. Researchers then identify text snippets that contain meaningful information and create *codes* (i.e. labels or tags) best describing the main insight of the snippet. After the researchers have refined a set of codes, they develop an initial description of the meaning of each code along with a *memo* – a short description explaining the code and elaborating on when it should be applied. Eventually, the codes from different evaluators are merged and discussed as a group. All codes as well as their memos are aggregated together into a code-book. The researchers then iteratively keep refining and re-evaluating the codebook until the process reaches saturation with a well established and shared understanding of all the codes.

The general inductive methodology involves five phases (Thomas, 2006). Ideally, this methodology results in the establishment of a hierarchical system of categories where codes are low-level components and categories are high-level generalizations of the codes. Every step along these phases has certain procedures associated with them. We now describe each of them, as well as the procedures we thereby undertook throughout our analysis:

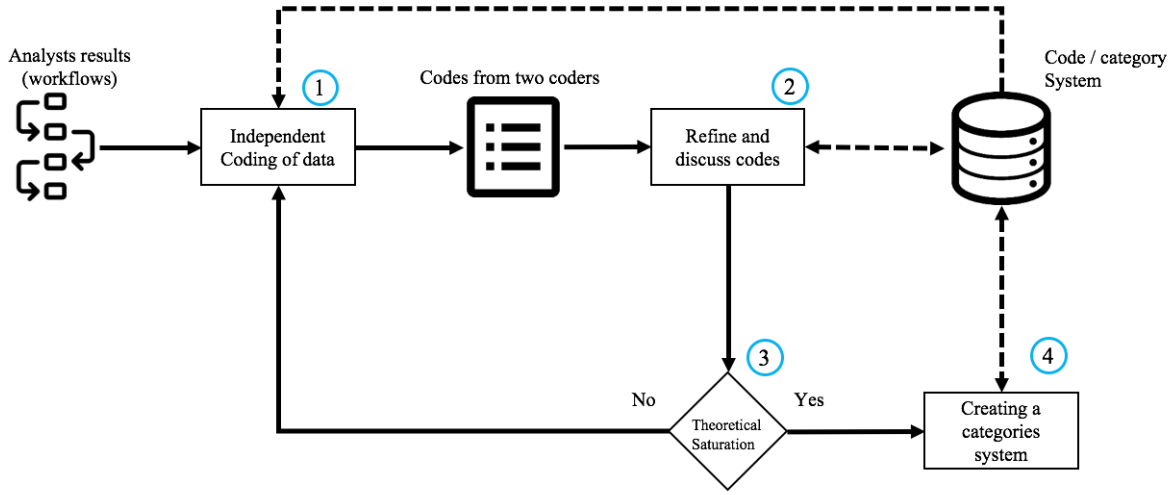

Figure 13. The workflow of our analysis

(1) Preparation of data (data cleaning):

The preliminary phase consists of transforming raw data into a common format and preparing the text for in-depth reading. In our case, we observed the procedure that data analysts followed throughout their work. Specifically, we recorded each of the commands executed and solicited their comments about the rationale of these steps. In analysis, oftentimes multiple commands address the same goal. For example, to analyze a dataset using linear regression, all variables have to be continuous. Hence, each categorical variable needs to be turned into dummy variables. The commands executed to transform all of those variables together represent one logical unit, which we call a block, with the goal of creating dummy variables in preprocessing to make the data amenable to linear regression.

To provide a useful unit of analysis, we enabled the participants of our study to split workflows (i.e. the whole sequence of all commands used in the analysis) into semantic blocks (essentially, sub-sequences of commands). This way, each block was annotated with descriptive properties which reflect the rationales and reasoning of the analyst's actions within a block (for a detailed description of a block, please refer to Figure 1). The structure of the descriptive properties originated in research on design rationale (Schubanz 2014) and design space analysis (MacLean et al. 1991). To summarize, the main goal of a block is to provide a unit of analysis with information about its purpose, reasoning and considered alternatives.

In our case there is no need for additional data preparation as the experiment is designed such that the aggregated data is already semi-structured with answers to

predefined questions about the goal and the considered alternatives in certain block. Since the data from each analyst is recorded in the same format, as advised by the analysis procedure proposed by Thomas, no further data cleaning is needed.

(2) Close reading of text (coding):

This first phase consists of detailed reading of the text until the researcher is familiar with the content and gains an understanding of major themes and concepts occurring in the text. In our study, the coders sequentially went through each block of an analyst's workflow and studied the descriptive properties. Following a simultaneous coding method (Step 1 in Figure 13), a coder can assign multiple codes to the same attribute of text (i.e. property of the block) (Saldana 2011). To help coders maintain consistent codes, we provided them with a searchable list of codes they had previously used. Coders could also retrieve all the explanations of the snippets annotated with the same code. This possibility encourages a coder to continuously compare the codes and refine her reasoning. A graphical workflow for the entire sequence of blocks, refined by the analysts at the end of their analysis, provides the coders with an overview of the relationship between the blocks. Embedded in the user interface, a coder can additionally assign explanations (i.e. short memos) for every coded text segment.

(3) Overlapping coding and uncoded text:

In our analysis, one attribute can be assigned to multiple codes and most of the text may not be relevant for the research. Moreover, coders did not have strict guidelines on what codes to propose (besides the general theoretical lenses of sensemaking we described in literature review), since the process is inductive and therefore not restricted. This way, while analyzing text snippets, different coders could apply codes of different granularity. To help understand the context of the code, the coders could additionally explain the codes assigned to the relevant text (Krippendorff 2004; Kurasaki 2000; Fahy 2001). As a result, the coded block might have different codes with a certain overlap, although the same key information was extracted by the coders.

Sometimes the answers provided by the analysts were not relevant for the research question. Hence, meaningless answers were not coded. Instead, coders were encouraged to apply codes that will explain *why* the analysts provided certain answers. We asked the coders to apply codes to block attributes in order to ease the task of interpreting. For example, the coders were asked to elicit the goal of the block, the considered alternatives, or why a certain alternative was preferred. Additionally, to capture the general purpose of a block, coders could also assign codes describing the general goal of the block.

(4) Creation of Categories:

In this phase coders collaboratively defined codes and discussed categories by summarizing and aggregating codes by their meaning (Step 2 in Figure 13). This was reached through a discussion where the meaning of the codes was clarified and the semantically identical codes have been merged. Further, based on the list of codes, the coders collaboratively constructed a category system – a high level organizing abstraction which summarizes the codes (Step 4 in Figure 13). Each refined category was provided with a memo, which summarized the coders thoughts and/or possible

relations to other codes/categories. These memos did not only served as justifications for the category, but also facilitated future revision and refinement of the category system. As Creswell (2002) suggests, this newly emerged list of categories should serve as new organizing scheme for coding. This is instrumental in inferring the categories based on the codes after a further iteration. In our case, the (updated) list of categories served as new coding scheme for coding in the subsequent iteration. This scheme is then to be applied to another subsample of the data, where coders can draw on the reasoning of memos when applying codes. Nevertheless, they can still come up with new codes, which are then either assigned to an existing category or build the foundation for a new category.

(5) Continuing revision and refinement of category system

Each iteration of coding ended with revision and refinement of the category system (Step 2 in Figure 13). The number of total assigned codes to a category as well as their prevalence could indicate the importance of categories. A category with only few assigned codes might indicate that this category is not well grounded. In such case, we considered merging this category with a more evident category (i.e. with more frequently occuring codes assigned to it) or removing this category. Miles et al. (2014) described the process of grouping initial codes into a smaller number of categories as *pattern coding*. At some point, we reached a theoretical saturation where no new codes and categories emerged from the new subsample of data (Step 3 in Figure 13). At this point, if the coders believe that each aspect accounting for the inherent variation in data analysis found in the data is captured in a category, the iterative coding is finished. A high percentage of agreement among the coders (i.e. proportional agreement) guaranteed not only a common understanding of the coding scheme, but also showed a high agreement when applying them.

All aforementioned procedures compose the main parts of the General Inductive Approach. There is another procedure relevant for the analysis – trustworthiness assessment. There are many ways to evaluate the trustworthiness for models developed in qualitative analysis. Literature proposes several ways to evaluate intercoder reliability or intercoder agreement, sometimes even contradicting each other(Campbell et al. 2013). According to Campbell *et al*., the use of such statistics for qualitative analyses aiming for systematic and rule-guided classification and retrieval of text are less imperative. As a consequence, simple proportion agreement (percentage of agreement among coders) is argued as reasonable approach(Kurasaki 2000). Moreover, some researchers claim, that looser standards are permissible in exploratory studies(e.g. Hruschka et al. 2004; Krippendorff 2004). In order to guarantee high reliability of the emerged final categories in this study, we applied both qualitative and quantitative measures of trustworthiness:

Independent parallel coding: Two coders independently developed a set of codes (step 1 in Figure 13). These two sets were compared and merged into a combined set (step 2 in Figure 13). When the overlap between the codes was low, the coders discussed and clarified each code in order to reach a more robust set of codes. This procedure also resembles the negotiated agreement approach proposed by Campbell *et al*., 2013.

Check on the clarity of categories: Two additional independent coders (previously not involved in coding) were presented to the set of codes supplemented with

explanation and examples. All coders were then asked to code a new subsample of data using the system of codes. Thereafter, If they came up with new codes the code-book was refined and translated into a new coding scheme. This coding scheme is then used for coding new data in another iterative cycle.

Calculation of interrater agreement: To measure the agreement among coders, we calculated the proportional agreement and Cohen's Kappa after each iteration (i.e. in Step 2 in Figure 13).

4. Study Design

To investigate our research question we designed an online platform called DataExplained that allowed participants to perform a data analysis online using RStudio. In the following subsections we described three major phases of the study we conducted in accordance with this study: i) the recruitment of participants, ii) the research setting, and iii) the analysis of submissions. More specifically, we first described how we recruited the analysts for our experiment and provided a short overview of the data. We then described the setting under which the data analysis was performed. Finally, the methodology of qualitative analysis of the results is explained.

4.1 Recruitment of Analysts

We approached potential participants via open calls on Twitter, Facebook, forums of psychology interest groups, platforms for collaboration and resource source exchange (i.e. StudySwap), and R mailing lists. The total duration of the crowdsourced project was about two and a half months during which the analysts would conduct data analysis using our platform.

In total, 132 people showed interest in participating in this crowdsourcing project, of which 47 carried out all steps involved in our study. The participants self-reported 7.34 years of experience in data analysis (SD=4.98) where 24 had a PhD, 13 Master, eight Bachelor and two a high school education with the average age of 30.95 (SD=6.29). Most of the analysts commented that they perform data analysis on a daily basis (20) or a few times a week (13), while the rest either perform data analysis once a week (5) or less (6). Most of the participants were male (38) and the rest are female (9). The majority of participants were residing in USA (25) whereas the rest are located in Europe (18) or somewhere else around the world. Besides ten participants who reported to not be currently associated with academia, the rest hold an academic position: eight are professors, six post-doctorants, fifteen are doctoral students and the rest hold another academic position (e.g., clinical psychologist, research analyst).

We next introduce the dataset and the analysis environment that analysts used in this study.

4.2 Dataset

The data analyzed in this study consists of eight thousand conversation threads taken from the academic forum Edge.com. As described by the Edge's founders, its purpose is to "To seek out the most complex and sophisticated minds, put them in a room together, and have them ask each other the questions they are asking

themselves.". We have constructed a dataset that allows for empirical tests of the role of a scientist's gender and status in intellectual conversations during 1996-2014 (the Edge currently has 797 contributors, of whom 140 are female). Data analysts were recruited using different social media platforms to maximize the number of scientists involved in the crowdsourcing experiment. The analysts were asked to test a set of key hypotheses derived from prior theory and evidence from the data. The first hypothesis is that woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion. sentence supporting this hypothesis from psychology literature(Ott 1989). The second hypothesis is that higher status participants are more verbose than are lower status participants. sentence supporting this hypothesis from psychology literature(Stein & Heller 1979).

4.3 The Methodology of Qualitative Analysis

Our analysis was guided by the evaluation objective of the study, thus focusing on the question of what factors are leading to the variability in data analysis. By doing so, we did not explicitly rely on any theory but let the findings arise directly from the interpretation of the raw data. We interpreted the blocks using all available information, such as the workflow of blocks and their description (taking into account future and past blocks), as well as analysts' comments and source code. Thus, our "coding filters" were broadly split into two areas. First, the objective output of a block, such as the method selection, the revision of code, or the task constraint. These factors are objective and all analysts face them equally throughout data analysis. Second is the subjective decision making process involved in data analysis. Factors such as personal beliefs, experiences, or intermediate insights which inform the next steps in a data analysis and differ among analysts. Thanks to the developed construct of "blocks" as well as the graphical representation of workflows, we had the necessary information to explain the rationale for every step in a data analysis.

First, two coders coded the blocks in a sequential manner, proceeding through blocks in their chronological order. For every applied code they provided an explanation to the code. As a result, every block was annotated with i) codes, ii) possible explanation, and iii) a reference to a relevant snippet, were it be analyst's verbal explanation or an executed code.

After both coders finished coding the blocks from a predefined subsample, the codes were grouped together. Next, the coders collaboratively refined and discussed each of them. As a result, similar codes were merged together, whereas too general codes were split into more self-explanatory codes. For each code, the coders created a short explanation in the form of a memo and provided some examples where this code has been applied. The resulting code-book (codes along with memo and examples) is then used for the subsequent coding iteration.

When theoretical saturation is reached, and the inter-rater agreement is high enough (proportional agreement and Cohen's kappa > 0.7), the code-book was presented to two additional coders. After they learned and refined the codes together with two initial coders, all four coders coded a new subsample and verified that the codes are suitable to describe the rationales perceived by the data analysts. In this phase, the code-book is further refined and new subsamples are coded until the agreement

among all four coders is high enough. In order to proceed to the next step, all coders iterated four times until the proportional agreement among them reached above 50%.

5. Results

Two coders followed three coding cycles in order to build a sustainable coding scheme. After each iteration, they discussed the discrepancies in the results and refined the codes. In the first iteration, both coders independently coded the 275 blocks of ten different analysts, to come up with an inclusive list of initial codes. As a result, they constructed 88 codes describing various factors of variability in data analysis. After eliminating duplicates (i.e. semantic synonyms) and not well-explained codes, they remained with 30 codes. To clarify whether these codes are inclusive and final, an additional subsample of 49 blocks corresponding of five analysts was then analyzed. During this iteration, coders realized, that some codes are too general and needed further refinement (i.e. either split the code in more detailed codes or delete the code entirely, as other codes may already substitute it). Therefore, they reviewed the blocks where rather high-level codes were applied and refined them to be more precise. The coding scheme for the last and third iteration consisted of 31 codes and 41 blocks. They then coded another subsample (21 blocks), however the code-book remained unchanged (i.e. they neither come up with new codes nor deleted any of already existing codes). The proportional agreement of the two raters after the last iteration was 72%, with a kappa measure of 0.7 (Cohen's Kappa). The resulting code-book was then presented to two new coders. They were provided with code-memos and examples of when (not) to apply each of the codes, and clarified any differences between related codes. All four coders then discussed the codes and clarified the codes with their corresponding memos. Following, the coders independently coded another subsample of 22 blocks. All four coders then discussed the results of their coding and updated the code book accordingly. The coders then coded again another subsample of 9 blocks. After the third iteration performed by all four coders, the percentage agreement reached 52.6%, and the code-book was finalized. Since there were no more disagreements at this point, there was no need for an additional coding iteration. At the end, the final code-book consisted of 31 codes (c.f. section 5.1). The four coders collaboratively grouped the codes together and created a category system with ten categories.

5.1 Codebook

In the following we describe the categories and list the corresponding codes. We provide a succinct explanation of the codes, a few examples of participants' corresponding comments, and a short discussion of how the categories contribute to the variability in data analysis. Some actions conducted by data analysts are not a result of one isolated consideration but rather a blend of multiple factors involved in decision making. Therefore, we often attributed multiple codes to a given analyst comment. In the examples provided here though, we always discuss every comment in the context of one, most descriptive) code.

Category	Codes	Category Description
Data	• **data constraint:** Any constraint imposed by the nature of data • **data quality:** Any objective metrics of data quality such as completeness, bias, distribution etc. • **feature engineering:** Adding new features (aka variables/columns /attributes etc.) which are a function of existing data. • **Preprocessing**: Any steps performed to preprocess the data (e.g. installing packages/libraries, removing outliers, organize data, etc.)	This category reflects all activities and considerations related to data. Data might have objective constraints such as format, missing values or size. Also, data transformation (i.e. feature engineering) and data preprocessing are data related activities which are not only changing the data but might lead channel data analysis in certain direction.
Examples of participant comments:		
1. "There's no variance in number of comments made. Data also has a temporal structure [so] that last analysis ignored" (**data constraint**) 2. "I don't think there is enough data to parameterize this model" (**data constraint**) 3. "Created paired changes in status normalized by the changes in the same period among people who did not change status" (**feature engineering**) 4. "Go through each row in original data, and only extract the first conversation of each thread" (**preprocessing**)		
Contribution to the variability of results:		
Data constraints limit and channel data analysis into certain direction. While sometimes these constraints can not be ignored (e.g. missing data, data size), it is a matter of expertise and experience to notice the problem in other cases. For example in (1), the analyst realized that the data is temporal. This made the results he/she had obtained invalid and resulted in a different approach being adapted instead.		

Another example is subjective decisions, such as what data to select as a subset of (4). In this case, the conclusions were derived based on this data. Would the analyst sample the data differently, and pick random conversations in each thread, the results could be different. Furthermore, analysts often transformed variables to be able to operate with more informative features (aka feature engineering). As it can be seen from (3), the way variables are transformed is a function of the analyst's internal hypothesis about the best way to operationalize the problem and may impact the further analysis.		
Task	• **task constraint:** Task constraint is related to the limitations imposed by the task analyst is performing (requirement by the task). For example, if the task is to report on certain measures or to produce a result up to certain deadline. • **complexity constraint:** Complexity constraint represents cases where analyst considers the complexity of alternatives or performed methods. A method might be objectively better but still avoided due to analyst's reluctance to engage in complicated data analysis process. This code is related to "effort constraint". However while the "complexity constraint" is related to the perceived complexity of the method (i.e. how complicated is it to execute), the effort constraint is related to the effort associated	**Task constraint** is related to the task which has to be accomplished during data analysis. This task could be either answering a hypothesis or an exploratory analysis aiming to produce potential research questions that could be answered with the data at hand. **Complexity constraint** represents cases where an analyst is considering the complexity of alternatives or performed methods. A method might be objectively better but still avoided due to the analyst's reluctance to engage in complicated data analysis processes. On the other hand, **task constraint** is related to the limitations imposed by the task the analyst is performing (i.e. task requirement). For example, when the task is to report on certain measures or to produce a result up to certain deadline.

	with alternative, which is not necessarily results from the complexity of the method. Another relevant code is a "methodological constraint". This code relates to the objective constraints imposed by the requirements of a method.	
Examples of participant comments:		
1. "not within scope of hypothesis" (**task constraint**) 2. "That the project requires the reporting of effect sizes and my approach - based on Bayes factors - cannot do that" (**task constraint**) 3. "complicated getting data into tm (date) format" (**complexity constraint**) 4. "More difficult to keep track of things" (**complexity constraint**)		
Contribution to the variability of results:		
When an analyst considers various alternatives of analyzing the data, task constraints and goals play a key role. For instance, if the task requires to report certain measures (2), or if the considered method requires the data to be in a certain format (3), the analyst will prefer certain analytical alternatives. Moreover, analysts might not proceed exploring some ad-hoc hypotheses that arose during analysis if they seem to be not within the scope of the task (1). Nevertheless, some of them could be helpful for answering the core questions of the overall analysis.		
Problem perception	• **uncertainty about the problem:** In this context by problem we mean the phenomenon which is under investigation. A problem analyst studies might be ambiguous in its nature due to different reasons. In addition, any uncertainties	This category refers to the problem the analyst is studying. This problem could be a hypothesis under investigation or an exploratory analysis. The perceived understanding of "problem mechanics" impacts an analyst's actions and informs intermediate steps throughout the data analysis. A problem an analyst studies might be ambiguous in its nature due to different reasons, such as loose specifications or different interpretations of certain aspects. In addition, any

	expressed with regards to the problem setting (e.g. if an analyst is not sure what is the meaning of variable in dataset, how the data was collected, or how to interpret the results) • **perceived understanding of the problem:** This code is applied when analyst is following a procedure due to the perceived logic of the problem. This code is mostly be applied, when a justification for the action is given with regards to the problem. Note, this code is different from the perceived understanding of reality. While perceived understanding of reality is reflecting a general context, understanding of the problem reflects a concrete problem analyst currently deals with and the sensemaking process that occurs. Selecting features/variables belongs to this code. • **intuition about the problem:** Intuition is a "gut feeling" that results out of prior knowledge or by inference from personal experiences, feelings and	uncertainties expressed with regards to the problem setting (e.g. if an analyst is not sure what the meaning of variable in dataset is, how the data was collected, or how to interpret the results) might affect the data analysis. Moreover, analysts often have an intuition about a problem. This kind of a "gut feeling" results from the prior knowledge or by inference from personal experiences, feelings and preferences. In data analysis, intuition might come into a play by unconsciously relying on it, in order to inform next intermediate analytical steps.

	preferences. Intuition in this case refers to intuitions about future actions.	
Examples of participant comments:		
1. "The scale is ordinal, but it's unclear to me how different each level is from the other - how much different is an experienced graduate student from a post-doc? An associate professor vs. a full professor? It seemed better to simply recognize them as nominal categories" (**uncertainty about the problem**) 2."It's hard to separate being female from many other factors that may also be the result of being female. Wanted to focus on a clean overall result without many controls. As noted in one alternative, couldn't come up with a reliable way to know if a female participant knew if there were other females in the conversation except for authors. There wasn't enough variation in number of times participating to use that to define active participation" (**uncertainty about the problem**) 3. "If hypothesis is that seeing women talk draws other women to be more active, the woman posting can only see that in regular discussions, not in annual conversations" (**perceived understanding of the problem**) 4. "These two variables atm seemed to be a good choice for the verbosity-operationalization, after going through all the language-variables created from the liwc-thingie" (**intuition about the problem**)		
Contribution to the variability of results:		
Research questions often hypothesize about high level artefacts. Operationalization of these artefacts is not always clear (1-2). This is where the analyst is mostly relying on the intuition about the problem. For example in (4), the analyst is stating that intuitively there are two variables in the dataset that might be a good representation of the verbosity artefact. Differently from intuition about a problem that can be seen in (4), in (3), the analyst expresses her perceived understanding about the problem. This means that there is much more certainty about understanding of the mechanics of the problem domain. For example in (3), there is a clear statement that the researched phenomenon can not be observed in certain data.		
Knowledge	• **perceived course of action:** The analyst performs an action in order to be able to continue the way she intends. (E.g. when analyst states a clear	The knowledge and experiences the analyst possess (e.g. when she refers to past analyses, claims to be familiar with a concept, or consequences of possible actions). The code "perceived course of action" describes a situation where the analyst performs a certain step in order to be able to further follow in a certain

	path to operationalize the problem - "Do A in order to do B"). • **personal knowledge:** Analyst's knowledge or prior experiences in performing an action she does (e.g. refers to past analyses, claims to be familiar with a concept, or consequences of possible actions) • **method preference:** Analyst's preference of certain methods. This can be either due to professional background/education or commonly faced problems. For example Bayesian statisticians prefer certain methods while some other researchers frequentist methods. • **expertise:** Decisions or actions that reflect professional knowledge and experience. For example when analyst is considering that while applying a certain method, one has to be careful of certain aspects such as assumptions or limitations. • **effort constraint:** Effort constraint represents cases where effort prevents analyst from taking certain actions/decisions	direction during the analysis. For example, when the data is transformed into a certain format in order to be able to apply an intended method (e.g. binarization of the outcome variable in order to perform a logistic regression). Furthermore, we observed expertise through decisions or actions that reflect professional knowledge and experience. For example, when an analyst is considering that when applying a certain method, one has to be careful of certain aspects such as assumptions or limitations. Awareness of the assumptions as well as consideration of methodological alternatives and their limitations, were seen as an indication of expertise. Last, effort constraint represents cases where effort prevents an analyst from taking certain actions/decisions during data analysis. This can be either due to time/complexity constraint or because the perceived benefit versus invested effort does not make it attractive (or "too much work to be done" as it was often reported).

	during data analysis. This can be either due to time/complexity constraint or because the perceived benefit versus invested effort do not make it attractive ("too much work to be done").	
Examples of participant comments:		
1. "Took data where each observation was a participant, and summarized it down to a dataset where each observation is a conversation. I wanted to be able to study things at the conversation level" (**perceived course of action**) 2. "these packages have been useful in my past analyses" (**personal knowledge**) 3. "Tried to run a Bayesian Hypothesis Test using the functions in BayesMed but it did not work" (**method preference/methodological constraint**) 4. "Models need to converge, and the choice of model terms cannot be data-driven since that would render the p-value for the chi^2 test meaningless due to the garden of forking paths" (**expertise**) 5. "More columns, harder to do tests based on blocks of variables" (**effort constraint**)		
Contribution to the variability of results:		
In (1) the analyst is summarizing the data to the conversation level in order to conduct further analysis on this level. Since the aggregation might often lead to information loss and lead the whole analysis in a certain direction, the perceived course of action contributes to the variability in data analysis. The same is true for the analyst's personal knowledge (2), method preference (3) and expertise (4), which all play a key role in predefining the course of data analysis. Lastly, the effort constraint is the factor that often undermines the depth of analysis. Like in (5), analysts often choose to avoid certain activities because they are time and effort intensive and will require too much of his or her resources.		
Belief	• **perceived understanding of reality:** The perceived understanding of the reality is a complementary factor to beliefs and	This category describes the tacit belief system of the analyst. Any personal assumption the analyst makes or action driven by personal interest of the analyst (e.g. curiosity or choices which relate to personal perceived rationales) might be categorized as part of the belief system. It is different from the explicit knowledge

	interests. Data analysts may have an implicit cognitive mechanism about "how things work" in the real world. This understanding is not directly about the problem which is under investigation but rather about a state in the grand scheme of things. • **personal assumption:** Any personal assumptions the analyst makes. For example, the analyst dropped most of the PhDs from my analysis as they will likely not influence the final result too much. • **personal interest:** Actions driven by personal interest of analyst (e.g. curiosity, choices which relate to personal rationales) • **personal preferences:** Analysts may have preferences or intentions to perform an action the way they think is best for them. These can be driven by various personal factors. If the preference is for a (statistical) method, we apply only the code "method preference".	by being tacit and unconscious by nature. It might be the personal belief (agenda) for an analyst to prove that a certain hypothesis is correct (e.g. the role of female in scientific discussions as they surfaced in one of the hypotheses we studied). Analysts may have preferences or intentions to perform an action the way they think is best for them. These can be driven by various personal factors. Such predisposition might play a key role in the way the data analysis is conducted even though no explicit traces can be observed in the data analysis results. The perceived understanding of the reality is a complementary factor to beliefs and interests. Data analysts may have an implicit cognitive mechanism about "how things work" in the real world. This understanding is not directly about the problem which is under investigation but rather about a state in the grand scheme of things.
Examples of participant comments:		

1. "I chose this option because there was no way to determine the value of job titles, however I think they are important. A director or a president has higher status than a graduate researcher and this should be reflected in the status" (**perceived understanding of reality**)

2. "Because the hypothesis is based on verbosity of users and not individual posts. My option assumes that total characters of each user is a strong metric for their overall verbosity" (**personal assumption**)

3. "interested in seeing how different disciplines have different gender breakdowns" (**personal interest**)

4. "Habits: I mostly start data analysis with such first steps" (**personal preferences**)

5. "I believe it's more robust" (**belief**)

Contribution to the variability of results:

Perceived understanding of reality describes the mental models of an analyst. For example in (1), once the analyst encountered an uncertainty, she relied on the perceived understanding of the importance of job titles. Hence, this variable was transformed into ordinal and included in the model. Other analysts would overlook this variable, and most likely even - if not - operationalize it differently (e.g. interpret the hierarchy of job titles). The analyst in (2) also makes a personal assumption while deciding to operationalise the verbosity through total number of characters. Additionally, when conducting an analysis, scientists are sometimes drifting from the core hypotheses in order to answer questions which are of their own interest, as exemplified in (3). The insights gained from this exploration inform the main analysis and have impact on the data analysis. Moreover, personal preferences and beliefs (4) inform the analysis and lead it in certain direction. For example, if one analyst starts her data analysis with data visualisation and exploration, the insights gained during this step might divert her from the initially anticipated course of analysis.

Exploratory data analysis	• **exploratory:** Any exploratory steps performed by the analyst. This is related to exploratory data analysis and can describe activities focused on data or model exploration. • **Visualisation:** Any kind of graphical visualisation / plot the analyst does. This is often related to the code "insight generation" or "exploratory analysis"	Exploring and understanding the data. This is related to exploratory data analysis and can describe activities focused on data or model exploration. For instance, data plotting and visualisation is often part of the exploratory data analysis where an analyst is attempting to understand data properties and their behaviour. This is also often related to the code "insight realization", since visualization often lead to new insights throughout the data analysis. Exploratory data analysis is well acknowledged as a cornerstone in data analysis (Tukey, 1977) and considered as a highly interpretative component that leads to decisions influencing the direction of the further analysis.
Examples of participant comments:		
1. “I experimented with both, but will ultimately use the non-transformed data for reporting; diagnostic plots did not improve much with transformations, and interpretability was reduced” (**exploratory**) 2. “Selected status metrics iteratively: identified several, plotted them, removed redundancies, plotted again and checked for correlations” (**exploratory**) 3. Looked at the univariate distributions for each column (Hmisc::describe()). Plotted the number of conversations/year. Plotted distribution of female participation as a density plot, then created scatter plots looking at male vs. female contributions; and # female contributors vs. female participation (**visualisation**)		
Contribution to the variability of results:		
Exploratory analysis is very common and occurring in many stages of data analysis. Even when the analysis is confirmatory by nature, analysts very rarely follow a predefined path to analyse the data. Mostly there is continuous exploration of the data that has impact on the way the analysis is conducted (aka adaptive data analysis) such as in (1-2). Visualisation (3) is one of the most powerful tools to explore data and is widely used throughout data analysis.		

confirmatory data analysis	• **revision of findings:** Revision of findings due to the *new* insights or idea. Often related to the code "insight realization" • **confirmatory measure:** Analyst tend to confirm their (intermediate) results in different phases of their analysis.	Reassures that the output makes sense and is correct. For example, that the data is indeed distributed according to the assumption, the results are within the expected range of values, or that the results are credible. Another example is a revision of findings due to the new insights or idea. Often the analyst has an insight or hypothesis about the problem and seeks to reconfirm it by checking whether the data corresponds to the anticipated behavior.
Examples of participant comments:		
1. "Ran the code from beginning to the end again, looked at the plots and rethought the modeling" (**revision of findings**) 2. "Re-ran the code to double check whether things are ok and to look in detail at the effect sizes and estimates" (**revision of findings**) 3. "Did a check with another analysis where I substituted Female with Male to reassure that the reversed coding of that variable didn't affect the R2" (**confirmatory measure**) 4. "Checked that the number of observations in the women-only subset was in line with what was expected" (**confirmatory measure**)		
Contribution to the variability of results:		
The category refers to reflection on the reached data analysis results. As stated by (1,3), often revision of the model sparks new insights and leads to remodeling steps. Experienced analysts often revise and compare the intermediate results in order to assure that the results are not flawed and make sense. This sensemaking process often leads to updating in the perceived understanding of the problem and causes analysts to reconsider the course of analysis.		

Methodology	• **uncertainty about the method:** If analyst is not sure whether the taken method is the correct one for her objectives or other method would fit better • ·**methodological constraint:** A methodological constraint related to the limitations imposed by considered methods or approaches. For example, assumption of normality or homoscedasticity have to be fulfilled in order to apply certain methods. • **interpretability constraint:** Analyst's have a subjective judgement for the interpretability of methods or approaches. This is a subjective constraint	Describes the methodological aspects of the conducted analysis. The methodological decisions might range from high level methodology to be used (e.g. Bayesian vs. Frequentist statistics) up to concrete decisions, such as how to operationalize the variables. Furthermore, analysts sometimes are not sure whether the selected method is the correct one for their objectives. Whenever we found an evidence for such uncertainty, we related this to the methodology. Lastly, a "methodological constraint" is related to the limitations imposed by considered methods or approaches. For example, assumption of normality or homoscedasticity have to be fulfilled in order to apply certain methods. Analysts have a subjective judgement for the interpretability of methods or approaches. This is a subjective constraint.
Examples of participant comments:		
1. "A mix of harder to model and not sure about the right assumptions" (**uncertainty about the method**) 2. "Unsure about whether I missed a covariate in the model and whether I need to change to a model accounting for the fact that the hierarchy variable is ordinal" (**uncertainty about the method**) 3. "Variables need to be at least ordinal", or, "model doesn't converge" (**methodological constraint**) 4. "This [method] seems simple, common-sense, and easy to interpret" (**interpretability constraint**)		

<table>
<tr><td colspan="3">Contribution to the variability of results:</td></tr>
<tr><td colspan="3">A decision of what method to apply is important and is often influenced by considerations, such as method sensitivity, robustness to assumption violations, and underlying approaches (e.g. Frequentist vs Bayesian). Additionally, when the method is hard to interpret (4) or mathematical modeling such that an alternative method could be applied is challenging (1-2), an analyst often opts for simpler, more transparent model. Hence, the uncertainty about alternative methods often results in analysts reusing the same, more familiar method across different datasets, even when they are aware of potentially more suitable methods. Since the statistical assumptions of methods are often open for discussion, analysts are often not sure how restrictive they should be with regards to this..</td></tr>
<tr><td>Insights</td><td>insight realization: This code describes a situation where the analyst generates new insights, instant hypotheses or ideas, due to the applied method/approach or throughout data analysis in general. This code can be seen as an evidence of sensemaking.
action driven by insight: Analyst's personal insights may drive certain actions to be followed (e.g. run correlation test on two variables of interest emerged from the insight generation). Often related with the code "Insight realization"</td><td>Reflects the insights gained throughout the data analysis. Insight realisation is a code that describes a situation where the analyst generates new insights, instant hypotheses or ideas, due to the applied method, approach or throughout the data analysis in general. This code can be seen as an evidence of sensemaking. Analysts' personal insights may drive certain actions to be followed (e.g. run correlation test on two variables of interest that emerged from the insight generation).</td></tr>
<tr><td colspan="3">Examples of participant comments:</td></tr>
</table>

<table>
<tr><td colspan="3">1. “I compared Threads to Job Title along with PhD Ranking, and found as prestige of Job Title increases, number of Threads increases, and this is especially true for higher PhD Ranks” (insight realization)

2. “Turned entries into paired data for people with word count and status. Thing repeated the process because I checked and realised sometimes people had more than one answer to an annual question”(action driven by insight)

3. “Prepared the individual entries for testing H2 based on the realisation that WC (i.e. word count) is sensitive to what year the communication was in” (action driven by insight)</td></tr>
<tr><td colspan="3">Contribution to the variability of results:</td></tr>
<tr><td colspan="3">One of the reasons for a data analysis to develop in a certain direction are intermediate insight realizations analysts have during the data analysis. For example, (2) had an insight, that, as prestige of Job Title increases, number of Threads increases. These realizations inform the decisions this analyst makes throughout her data analysis. For example the insight (3) had about “WC (i.e. word count) is sensitive to what year”, triggered the restructuring of data, in order to better account for this phenomenon.</td></tr>
<tr><td>Coding skills</td><td>• code quality: Actions performed to enhance the objective quality of code (e.g. reorganize, refactor, comment etc.)
• debugging: Code executed for debugging / corrective measures.</td><td>Since we explore a case where the data analysis is conducted without a user interface mediation but through R coding, these actions describe coding skills of an analyst. Code quality relates to the measures undertaken by an analyst to enhance the objective quality of code (e.g. reorganize, refactor, comment etc.). Debugging code relates to the activities whose purpose is to find an error in code that presents unintended or wrong results. It also includes activities related to test the correction.</td></tr>
<tr><td colspan="3">Examples of participants comments:</td></tr>
</table>

1. It's cleaner code since I only use it for a few variables (**code quality**) 2. Rewrote and commented the code (final pretty version.R) so that it was better for sharing, then reran the analysis of the code (**code quality**) 3. Caught an error, rerunning analysis with error fixed (**debugging**) 4. Troubleshooting the aggregation by participant (**debugging**)
Contribution to the variability of results:
The major contribution of code quality to the variability in data analysis is through the complexity that it introduces. When the code is not well organised and confusing this impacts data analysts to further explore more complicated alternatives (1). On the other hand, the code quality is correlated with the expertise of the analyst. This means that non-experts are less prone to explore alternatives and rather stick to simple to follow (and code) approaches.

5.2 Organizing model

In line with the rationale design approach, we further grouped the categories into four major meta-categories that explain different considerations which can drive an analytical approach (marked yellow in Figure 5). These groups represent meta-categories based on their function in the model of cognitive processes involved in data analysis we propose here. In the following we describe each of them and outline their function on a broader scheme of variability of data analysis.

What (specifications): This characteristic entails the categories which are (a priori) given and objective in nature (i.e. the same for different data analysts). The categories belonging to this characteristic are *Data* and *Task*. Note that these factors might still be interpreted in various ways (e.g. due to new insights or personal beliefs), but cannot be changed. These categories serve as a starting point of the data analysis and a point out where objective (i.e. given) specifications meet subjective personality of a data analyst. Having data and task (e.g. hypothesis to test) at hand, the analyst then proceeds to understand the data. This is where the first source of variability can be observed due to the personal differences among data analysts.

Who (personal): The second characteristic relates to all personal attributes of a data analyst. This characteristic includes the categories *Knowledge*, *Belief,* and *Problem perception* which reflect the contribution of the personal biases and attitudes in problem-solving in general as well as in data analysis. The differences in data interpretation lead to different activities of preprocessing and collection of additional data. For example if the data does not support the current understanding of the problem, an analyst might be prone to collect additional data that will support her beliefs. However, it is much more common that the way data is preprocessed (cleaned, subsampled, aggregated etc.) is a consequence of personal factors, leading to a variability.

The interplay between the first two meta-categories is also referred by Grolemund and Wickham (2014) as interaction between mental models and given data. Throughout the process of studying and understanding the data, an analyst updates her prior beliefs and biases with regards to what was expected vs. what is actually reflected in the data. Sometimes these discrepancies lead to updated beliefs, while in some cases an analyst internally offers an alternative explanation for the observed mismatch and rejects an alternative state of belief. This process is to some extent similar to the statistical hypothesis testing where the alternative hypothesis is either accepted or rejected. The difference is that in this case it occurs in the analyst's mind and the process is not well understood. An example for this could be a certain (perceived) understanding resulting from the professional background or personal experiences which is challenged by the data which does not support this with evidence and, therefore calls these a-priori understandings into question.

How (analysis): "How" relates to categories that accounts for actions or methods which are performed during data analysis. This meta-category is a confounder to the variability of data analysis, since the way data analysis is carried out influences the final results. The categories belonging to "how" either have an exploratory or confirmatory character. The methods chosen to achieve the desired results thereby vary among different analysts. At some points during the data analysis, an analyst might reach insights which interact with the personal understanding of the problem and the system of beliefs (i.e. cognitive sensemaking process). We can thereby distinguish between two general data analysis approaches in this context: We refer to exploratory data analysis (EDA) as the process of data exploration, as well as attempts to understand the logic of the problem and summarizing its main characteristics. Confirmatory data analysis (CDA) refers to the analytic choices to confirm the emerged models (i.e. systematically assess the strength of evidence) in an iterative way [Hoaglin, 2003]. As an example, assume that an analyst wants to find out the relation between two variables of interest. She therefore applies different methods (e.g. runs a correlation or plots different diagrams), in order to understand this relationship on a subset of the data (EDA). Once the analyst seems to have understood the meaning of these variables (i.e. made sense of the data/problem), she wants to confirm her insights and fits a linear model on another subset of the data (CDA). This interplay between exploration and confirmation can be observed in various stages of a data analysis, since the insights of CDA may not necessarily be in line with the findings of the exploratory phase. In that case, further exploratory steps might be necessary. With multiple iterations of EDA and CDA, the analysts continuously refine their analysis. This cycle ends, once the analyst reports her final findings with regard to the stated hypotheses.

Where (sensemaking): data analysis is an iterative process that can be seen as a spiralic process where each iteration leads to new insights gained. As a result, an analyst makes decisions on how to proceed with her data analysis, and advances further in a certain direction. The "Where" meta-category is the starting point of each such iteration where the analysts process the results of the previous iteration and make a decision on how to proceed. During this process, an analyst decides whether to confirm, update or reject the current understanding of the problem due to insights gained from the previous iteration. Even though it is unclear what the mechanism leading to each one of this outputs is, it is apparent that they play a

significant role in deciding about the next steps to be followed by the analyst throughout her data analysis. The world-view (i.e. the understanding of reality and mental schemata) helps analysts determine where to allocate more attention and how to interpret the the data (Klein & Moon 2006). Information that does not match the world-view is likely to be either overlooked, explained with alternative theory, or updated if the signal coming from the data is significantly strong.

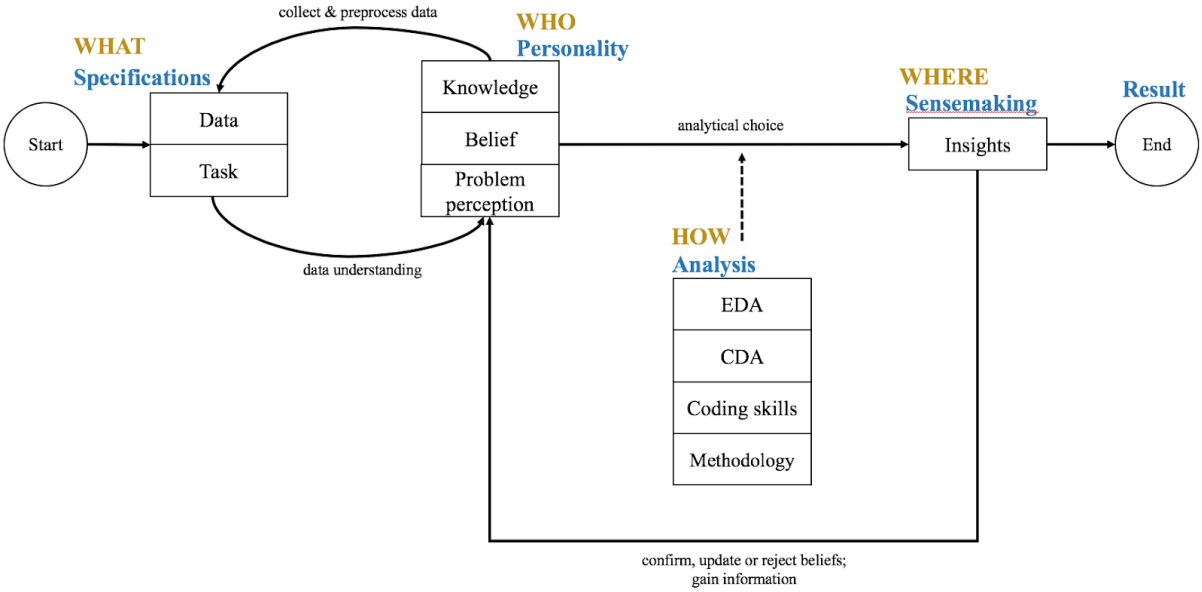

Figure 14. Categories in a data analyst's workflow

5.3 Illustration of the variability in data analysis: comparison of two workflows

In the following we demonstrate how the proposed model (Figure 14) reflects the data analysis workflow of two data analysts who took part in our study. Each one of them followed the workflow proposed in Figure 14 and ended up with very different results. We walk through the decisions made during their analysis and discuss how decisions made at each step could (and probably did) bias the results.

The analysts received a dataset along with the task description that included testing two hypotheses. In this example we only demonstrate the analysis of hypothesis 1: *"A woman's tendency to participate actively in the conversation correlates positively with the number of females in the discussion"*.

Our first analyst is a 37 year-old female, which has a PhD degree in political science with extensive statistical training. She holds a Masters in statistics from a top university and has fifteen years of experience in data analysis. The second analyst is a 43 year-old male, professor with a PhD degree in behavioral sciences with extensive statistical training and experience of teaching multiple statistics classes for graduates. The analyst has fifteen years of experience of statistical analysis with R.

Both analysts have an academic background with impressive records in data analysis. Both perfectly fit into a profile of researchers publishing data-driven research papers. And yet their results are very different. The result of the first investigator suggest the effect size of odds ratios to be within the confidence interval

between [16.2, 236.1], while the results of the effect size of odds ratios of the second investigator is with confidence interval [1.05,1.07]. Even though both conducted legitimate data analysis they ended up with different results. We argue that the difference is a result of the process outlined in Figure 5. Next we go through the analysis of both analysts and demonstrate how the output of their analysis was preconditioned on the decisions they made throughout data analysis.

5.3.1 First analyst: Data analysis description

After exploring what variables are present in the dataset, analyst started exploring the variables that in her opinion were associated with female participation in discussions. The next step was to exclude the discussions without or with only one woman involved. **This is the first time we see a subjective decision of the analyst based on the problem perception.** Would this data not be excluded it could lead to biased result depending on the method to be applied in the later steps of the analysis. Furthermore, the analyst applied a confirmatory measure in order to verify that the data is consistent: every thread ID (i.e. discussion on Edge.org) is matching only one web link. As a result, a few non consistent threads were identified and updated. **This action is a result of personal knowledge and experience of analyst and was triggered by the insight that the relationship between the provided thread ID and the web link is not necessarily of high quality.** The next activity resulted from the understanding that the new variable that summarises the number of authors for every thread is necessary. This can be attributed to the feedback loop between "Who" and "What" were the data is updated according to what is necessary to proceed with the data analysis. Analyst perceived this information as necessary since her intermediate goal was to explore the distribution of number of female authors per thread (there might be more than one author) and compare it with the number of the unique female contributors for every thread. After getting this insight, she decided to focus only on conversations with more than one author and only those which were not live discussions. This exploration led to new insights: two more threads that were incomplete were found, and they were excluded from data analysis as a consequence. **Again, the exclusion of data is a function of the interaction between the specifications of the analysis (i.e. task and data) and the personal attributes of the data analyst.** Then, as part of data exploration, the count of threads each user participated in was counted. As mentioned by the analyst, this was done out of interest and as part of data exploration. **This is an example for how personal interest can be present in data analysis.** Another iteration of data exploration was the creation of a feature that sums up the number of contributions for every person in every thread, of both author and commentator type. This feature engineering step was done in order to be able to explore the number of contributions for every person as author and commentator grouped by thread and gender, which is core of the hypotheses under analysis. **Latter is an example of an action driven by insights, where the analyst's perception of the problem under investigation leads to a certain sequence of activities, ending up with the information the analyst perceives as helpful or necessary to answer the research question.** Then, analyst decides to operationalise "active participation" as threads with comments. As a result, she considered to exclude all threads without comments. **The problem**

perception in this case resulted in two activities that might route the analysis: the way "active participation" is operationalised, and, as a result, the exclusion of threads without comments when they would be found. Next, the analyst intended to create a new variable, which corresponds to preprocessing the data and more specifically to feature engineering. She however realised, that the originally loaded data is of bad quality, having some rows omitted. Would the analyst not realise this problem, the issue of data quality would have an impact, as the results could be biased. **This is an example of how insight realisation can lead to corrective activities of debugging and code quality improvements, such that no data is omitted when loaded.** This also corresponds to the knowledge and experience of the data analyst to be able to spot data anomalies and correct them on time.

The next activity is related to the problem perception, knowledge, and also, to some extent, to the belief system. The analyst is checking whether there is a need to control for authors of their own comments or not, in order to avoid bias in the data. To do so, analyst looked at the number of times a female author is also listed as a contributor. Afterwards, more exploratory data analysis was conducted, in order to visually observe how the proportion of female commentators in threads without female authors is different from threads with female authors. After observing the result, analyst reached an insight that the difference might be limited to certain areas in science and she therefore computed proportions based on author discipline. **This is an example of a sensemaking process where sparked insights lead to a new cycle of reviewing the current beliefs and gaining new knowledge.** Finally, analyst computed the odds ratio for female commentator when female author versus male commentator. Another considered alternative was to calculate the odds ratio for every discipline by controlling for differences across different fields, like the number of women being active overall in each field. However, the data type made this calculation harder and the results across disciplines seemed similar. As analyst said, "there are 0s that prevent calculation of odds ratios by discipline; while there is some variation across disciplines, the result doesn't appear to be isolated to particular disciplines, so keeping them combined gives a more simple result". **In this case, the effort constraint as well as data constraint representing the interplay between the specifications and the personality of the analyst played a role in not proceeding with this alternative**. Another considered alternative was regarding different operationalisation strategies of active participation. Namely by looking at the overall number of females in a thread rather than female commentators as a function of female author. However, analyst considered this operationalization as not informative enough, even though it would be more in line with the hypothesis. Guided by a **personal belief and problem understanding** this alternative was avoided.

Then, analyst conducted activities associated with **confirmatory analysis**, in order to make sure that the result also holds when controlling for disciplines and threads with more than one author. Analyst realised that she was including cases where there was no commentator, and consequently recomputed odds ratios using threads with at least one commentator. The analysis concluded with a confirmatory analysis involving data plotting and verifying that the results are robust and make sense. The results reached in this analysis are odds ratios of 52.3 within the confidence interval of [16.2, 236.1].

Now, we want to compare the analysis of the first analyst to the second.

5.3.2 Second analyst: Data analysis description

After loading the data the analyst is exploring the data by reducing the data to include only few features. As analyst explained it *"I couldn't get a grip of the data so I had to make it smaller (reduce the number of features) to get an overview"*. **This is an example of how the data constraint (i.e. data is too big to grasp) intertwined with personal constraint to interpret the data in initial state lead to a decision to reduce the data.** Moreover, the analyst decides to leave certain features only: Thread ID, the percentage of unique female participants, the unique identifier of the contributor, boolean variable that signifies whether the commentator is female and the order of the text pieces. A given feature selection represents perceived understanding of the problem and fits into a problem perception category belonging to the personality of the analyst. Note, that different subset of features could lead to different subsequent analysis.

After the reduced data was observed, the analyst had an hypothesis that every female contributes only once to every thread (i.e. Female_Contributions = UniqueFemaleContributors). To test this hypothesis analyst plotted these two variables. **This hypothesis is a result of sensemaking that occurred as a result of exploratory steps and as a function of analysts problem perception.** The reason why this is important for analyst is explained by him due to the fact that it is easier to do the operations of creating new variables, which hints to personal knowledge of participant being a driving factor in this decision. The way analyst evaluated whether two variables are identical enough through plotting is also referred as effort constraint by the analyst. Alternative way would be by applying statistical test that would determine whether these two variables are significantly different. After the author is convinced that the variables are the same he constructs two new variables of number of female contributors in discussion and the next female contributor in the discussion. These two variables are a consequence of the insight gained in the previous step and part of the perceived course of action conducted by the analyst. Next, the analyst plotted these two variables in order to inform his next step. Finally, the analyst is regressing the next_female variable as a function of number of females in discussion, which is the another variable that he previously created. The analyst also considered multilevel analysis that would take into account other variables but **preferred logistic regression due to better acquaintance with this method**. Moreover, the analyst mentioned that he does not understand the text analysis research field well enough to do it properly and therefore opted for a more conventional approach. These comments mark factors like effort constraint and method preference as driving factors for the chosen course of analysis. Eventually the analyst reported the confidence interval of effect size of odds ratio to be between [1.05,1.07].

In the examples above we demonstrated how the analysis conducted by two expert data analysts resulted in very different results. These highlights the problem inherited in data driven analysis: the existence of multiple factors driving subjectivite decisions which are made throughout data analysis. Next we discuss the implications of these factors and propose how they could be made transparent and agreed upon.

6. Discussion

Our study is in line with previous research demonstrating the subjective nature of data analysis. Even when provided with the same dataset and predefined hypotheses, researchers often reach varying conclusions due to different operationalization of variables, data analysis strategies, and personal beliefs and constraints. Therefore, by means of the DataExplained platform and subsequent qualitative analysis, we explored which factors caused this variability in results. We also proposed a model helping to describe the process through which analysts reached their conclusions. This model draws on sensemaking theory and extends the conceptual model of data analysis proposed by Grolemund and Wickham (2014) by outlining how personal factors and task specifications interact to drive variability in outcomes (Figure 5). The proposed model was empirically derived and, to the best of our knowledge, is the first study to provide a detailed, data grounded overview of the behavioral factors involved in the data analysis process lead to variability in the results.

The results of our study inform the discussion regarding the ongoing crisis of confidence in science. The accumulating evidence from this project and others suggests that ostensibly data driven findings are subject to concerns about their robustness and reliability. Researchers must make subjective choices regarding how they obtain, aggregate, clean, model, and interpret data. As a result, many findings may not be robust to different defensible operationalization of variables and analytical choices made by researchers. This may be especially true when dealing with controversial issues about which different scientists may have different priors, and for complex data sets in which a variety of defensible analytic approaches could be adapted. Therefore, there is a need for greater humility and caution in interpreting findings from data driven research based on complex datasets and when the researched phenomenon is not yet well understood. Moreover, not only academics but also practitioners would benefit from caution when considering acting on on scientific discoveries from a single research team. Rather than rely on a single analytic team or report, organizations should have multiple (smaller) teams analyze the same data and compare results before making major strategic decisions. In this research we used crowdsourcing to uncover which factors drive variability in results. Crowdsourcing may be especially useful for controversial topics with public policy implications where research transparency is most criticalOne the other hand, crowdsourcing data analysis is not feasible for all projects, and integrating DataExplained or similar platforms into single-team research projects can help make transparent the role of subjective choices in the reported results. Some researchers might be reluctant to justify and describe all decisions during data analysis. In the long term it might be possible to reduce the workload of logging reasons for analytic decisions made by different analysis teams with the help of Artificial Intelligence, such that a single planner develops multiple analysis plans with the assistance of AI. In such case auto-experimentation may reduce the workload and prune certain approaches.

The results of this crowdsourced initiative demonstrate a broader problem for science than selecting an analytic approach to get significance, or peeking at the data and then testing for what look like significant relationships, both of which can be addressed via mandatory pre-registration. The phenomenon of analysis contingent

results, such that there exists a broad range of defensible-but-subjective decisions that impact the research conclusions, will not be eliminated by committing to the analysis plan beforehand. However a crowdsourced approach in which the analysis process is made transparent and every decision point can be identified and deliberated might serve as a potential way to address this challenge.

7. References

Affairs, L.W. and the T.F. on S.I. a P. a B. of S., 1999. Statistical methods in psychology journals. *American psychologist*, 54 (8)(8), pp.594–604.

Agrawal, A. et al., 2013. Digitization and the Contract Labor Market: A Research Agenda. *NBER Working Paper*, p.37.

Alasuutari, P., 2010. The rise and relevance of qualitative research. *International Journal of Social Research Methodology*, 13(2), pp.139–155.

Anderson, C., 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Magazine*, 16(7), pp.1–2. Available at: http://www.wired.com/print/science/discoveries/magazine/16-07/pb_theory.

Baker, M., 2016a. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), pp.452–454. Available at: http://www.nature.com/doifinder/10.1038/533452a.

Baker, M., 2016b. Statisticians issue warning on Pvalues. *Nature*, 531, p.151.

BARNES, W.H.F., 1944. The Nature of Explanation. *Nature*, 153(3890), pp.605–605. Available at: http://www.nature.com/articles/153605a0.

Bernstein, A., 2000. How can cooperative work tools support dynamic group process? Bridging the specificity frontier. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. pp. 279–288.

Bernstein, A., Klein, M. & Malone, T.W., 2012. Programming the global brain. *Communications of the ACM*, 55(5), p.41.

Bernstein, M.S. et al., 2010. Soylent: a word processor with a crowd inside. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp.313–322.

de Boer, P.M. & Bernstein, A., 2015. PPLib: Towards the Automated Generation of Crowd Computing Programs using Process Recombination and Auto-Experimentation. *ACM Transactions on Intelligent Systems and Technology*, (Special Issue: Crowd Computing).

Bollier, D., 2010. *The Promise and Peril of Big Data*, Available at: http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf.

Boyd, D. & Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, 15(5), pp.662–679.

Campbell, A. & Wu, A.S., 2011. Multi-agent role allocation: Issues, approaches, and multiple perspectives. *Autonomous Agents and Multi-Agent Systems*, 22(2), pp.317–355.

Campbell, J.L. et al., 2013. Coding In-depth Semistructured Interviews: Problems of

Unitization and Intercoder Reliability and Agreement. *Sociological Methods and Research*, 42(3), pp.294–320.

Carpenter, J., 2011. May the best analyst win. *Science (New York, N.Y.)*, 331(6018), pp.698–699.

Chi, M.T.H., 2008. Three Types of Conceptual Change: Belief Revision, Mental Model Transformation, and Categorical Shift. In *Handbook of research on conceptual change*. pp. 61–82.

Collaboration, O.S., 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251), p.aac4716-aac4716. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716.

Conklin, E.J. & Yakemovic, K.C.B., 1991. A Process-Oriented Approach to Design Rationale. *Human-Computer Interaction*, 6, pp.357–391.

Creswell, J., 2002. *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*,

Davenport, T.H. & Patil, D.J., 2012. Data_Scientist-the_Sexiest_Job_of_the_21St_Century.Pdf. , pp.70–76.

Davey Smith, G. & Ebrahim, S., 2002. Data dredging, bias, or confounding. *Bmj*, 325(7378), pp.1437–1438. Available at: http://www.bmj.com/cgi/doi/10.1136/bmj.325.7378.1437.

Van Dijck, J. & Nieborg, D., 2009. Wikinomics and its discontents: a critical analysis of Web 2.0 business manifestos. *New Media & Society*, 11(5), pp.855–874. Available at: http://journals.sagepub.com/doi/10.1177/1461444809105356.

Dissanayake, I., Zhang, J. & Gu, B., 2014. Virtual Team Performance in Crowdsourcing Contests : A Social Network Perspective. *ICIS 2015 Proceedings*, (Savage 2012), pp.1–16.

Dwork, C. et al., 2015. validity in adaptive data analysis. *Science*, 349(6248), pp.636–638. Available at: http://www.sciencemag.org/content/349/6248/636.

Edgell, S.E. & Noon, S.M., 1984. Effect of violation of normality on the t test of the correlation coefficient. *Psychological Bulletin*, 95(3), pp.576–583.

Eicken, H., 2013. Six red flags for suspect work. *Nature*, 497, pp.433--434.

Erceg-Hurn, D.M. & Mirosevich, V.M., 2008. Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7), pp.591–601. Available at: http://doi.apa.org/getdoi.cfm?doi=10.1037/0003-066X.63.7.591.

Fahy, P., 2001. Addressing some common problems in transcript analysis. *The International Review of Research in Open and Distance Learning*, 1(2), pp.1–6. Available at: http://www.irrodl.org/index.php/irrodl/article/viewArticle/321.

Feldman, M., Juldaschewa, F. & Bernstein, A., 2017. Data Analytics on Online Labor Markets: Opportunities and Challenges. Available at: http://arxiv.org/abs/1707.01790 [Accessed August 12, 2017].

Feldman, Mi., Anastasiu, C. & Bernstein, M., 2016. Towards Enabling Crowdsourced Collaborative Data Analysis. *Collective Intelligence*, (June), pp.1–5.

Fernandes-Taylor, S. et al., 2011. Common statistical and research design problems in manuscripts submitted to high-impact medical journals. *BMC research notes*, 4(1), p.304. Available at: http://www.biomedcentral.com/1756-0500/4/304 [Accessed July 29, 2015].

Field, A., 2013. Discovering Statistics using IBM SPSS Statistics. *Discovering Statistics using IBM SPSS Statistics*, pp.297–321.

Fiske, S.T., 2016. How to publish rigorous experiments in the 21st century. *Journal of Experimental Social Psychology*, 66, pp.4–6. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0022103116000032.

Fox, P. & Hendler, J., 2011. Changing the equation on scientific data visualization. *Science*, 331(6018), pp.705–708.

Friedkin, N.E. et al., 2016. Network science on belief system dynamics under logic constraints. *Science*, 354(6310), pp.321–326.

Gelman, A. & Hennig, C., 2015. Beyond subjective and objective in statistics. *arXiv preprint arXiv:1508.05453*. Available at: http://arxiv.org/abs/1508.05453 [Accessed September 16, 2016].

Gelman, A. & Loken, E., 2014a. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Psychological bulletin*, 140(5), pp.1272–1280. Available at: http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf%5Cnhttp://doi.apa.org/getdoi.cfm?doi=10.1037/a0037714.

Gelman, A. & Loken, E., 2014b. The statistical Crisis in science. *American Scientist*, 102(6), pp.460–465.

Gelman, A. & Shalizi, C.R., 2015. Philosophy and the practice of Bayesian statistics Andrew. *British Journal of Mathematical and Statistical Psychology*, 66(1), pp.8–38.

Gilad-Bachrach, R., Navot, A. & Tishby, N., 2004. Margin based feature selection - theory and algorithms. In *Proceedings of the 21st International Confer- ence on Machine Learning*. p. 43. Available at: http://eprints.pascal-network.org/archive/00000869/.

Glaser, B.G. & Strauss, A.L., 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*, Available at: http://www.amazon.com/dp/0202302601.

Glass, G. V, Peckham, P.D. & Sanders, J.R., 2012. Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance Author (s): Gene V . Glass , Percy D . Peckham and James R . Sanders Reviewed work (s): Source : Review of Educational Research , Vol . 42 , N. *Review of Educational Research*, 42(3), pp.237–288.

Good, P.I. & Hardin, J.W., 2012. *Common errors in statistics (and how to avoid them)*, John Wiley & Sons.

Gregor, S., 2006. The nature of theory in information systems. *MIS Quartely*, 30(3), pp.611–642.

Grolemund, G. & Wickham, H., 2014. A Cognitive Interpretation of Data Analysis. *International Journal of Statistics*, 82(2), pp.184–204. Available at: http://vita.had.co.nz/papers/sensemaking.pdf%5Cnhttp://onlinelibrary.wiley.com/doi/10.1111/insr.12028/abstract.

Gruber, T.R. & Russell, D.M., 1993. Generative Design Rationale: Beyond the Record and Replay Paradigm. *Design rationale: Concepts*, (December 1993). Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.1981&rep=rep1&type=pdf%5Cnpapers2://publication/uuid/AEA45CFD-89DF-4DE4-BC2D-71283E2E5DFB.

Guindon, R., 1990. Knowledge exploited by experts during software system design. *International Journal of Man-Machine Studies*, 33(3), pp.279–304.

Gutierrez, D.D., 2015. *Machine learning and data science: an introduction to statistical learning methods with R*, echnics Publications.

Haas, D. et al., 2015. Wisteria : Nurturing Scalable Data Cleaning Infrastructure. *Proceedings of the 41st International Conference on Very Large Data Bases*, 8(12), pp.2004–2007.

Head, M.L. et al., 2015. The Extent and Consequences of P-Hacking in Science. *PLoS Biology*, 13(3).

Heer, J., Viégas, F.B. & Wattenberg, M., 2009. Voyagers and Voyeurs: Supporting Asynchronous Collaborative Visualization. *Communications of the ACM*, 52(1), pp.87–97.

Hevner, A.R. et al., 2004. Design Science in Information Systems Research. *MIS quarterly*, 28(1), pp.75–105.

Hill, R.C. & Levenhagen, M., 1995. Metaphors and Mental Models: Sensemaking and Sensegiving in Innovative and Entrepreneurial Activities. *Journal of Management*, 21(6), pp.1057–1074.

Hoekstra, R., Kiers, H.A.L. & Johnson, A., 2012. Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3(MAY), pp.1–9.

Howison, J. & Crowston, K., 2013. Olla boration through open superposition.

Hruschka, D.J. et al., 2004. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods*, 16(3), pp.307–331. Available at: http://journals.sagepub.com/doi/10.1177/1525822X04266540.

Humphreys, M., Sanchez de la sierra, R. & Van der windt, P., 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. *Political Analysis*, 21(1), pp.1–20.

Intelligence, A., 1992. Task-Structure Analysis Knowledge MOdeling for. , 35(9).

Introne, J. et al., 2013. Solving wicked social problems with socio-computational systems. *Kuntsliche Intelligenz*, 27(1), pp.45–52. Available at: http://cci.mit.edu/working_papers_2012_2013/cciwp2012-05colabkunstinel.pdf.

Ioannidis, J.P.A., 2005. Why most published research findings are false. *PLoS Medicine*, 2(8), pp.0696–0701.

Johnson-Laird, P.N., 1980. Mental models in cognitive science. *Cognitive Science*, 4(1), pp.71–115.

Jussim, L. et al., 2015. Interpretations and methods: Towards a more effectively self-correcting social psychology ☆. *Journal of Experimental Social Psychology*, xxx, pp.116–133. Available at: http://dx.doi.org/10.1016/j.jesp.2015.10.003.

Kalleberg, A.L. & Dunn, M., 2016. Good Jobs, Bad Jobs in the Gig Economy. *The Gig Economy: Employment Implications: Perspectives on Work 2016*, 20, pp.10–14.

Kandel, S. et al., 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. *Human factors in computing systems. ACM*, pp.3363–3372.

Kanji, G. k, 2006. *100 Statistical Tests* 3rd ed., London: SAGE Publications India Pvt Ltd. Available at: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006199-199501000-00015.

Kaptein, M. & Robertson, J., 2012. Rethinking Statistical Analysis Methods for CHI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.1105–1114. Available at: http://doi.acm.org/10.1145/2207676.2208557.

Kay, M., Nelson, G.L. & Hekler, E.B., 2016. Researcher-Centered Design of Statistics. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, (August), pp.4521–4532. Available at: http://dl.acm.org/citation.cfm?doid=2858036.2858465.

Kittur, A. et al., 2011. CrowdForge: Crowdsourcing Complex Work. *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11*, pp.43–52. Available at: http://dl.acm.org/citation.cfm?doid=2047196.2047202.

Kittur, A. et al., 2012. CrowdWeaver : Visually Managing Complex Crowd Work. *Scenario*, pp.1033–1036. Available at: http://www.cs.cmu.edu/~pandre/pubs/crowdweaver-cscw2012.pdf.

Kittur, A., Nickerson, J. & Bernstein, M., 2013. The Future of Crowd Work. *Proc. CSCW '13*, pp.1–17. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2190946%5Cnpapers2://publication/uuid/AE6BF263-1DEF-4900-8C95-DC8BAD2DE4AF.

Klein, G. & Moon, B., 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, 21(5), pp.88–92.

Klein, R.A. et al., 2014. Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), pp.142–152.

Krippendorff, K., 2004. *Content Analysis: An Introduction to Its Methodology*,

Krishnan, S. et al., 2015. SampleClean: Fast and Reliable Analytics on Dirty Data. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, pp.59–75. Available at: http://sites.computer.org/debull/A15sept/p59.pdf.

Kulkarni, A., Can, M. & Hartmann, B., 2012. Collaboratively crowdsourcing workflows with turkomatic. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*, p.1003. Available at: http://dl.acm.org/citation.cfm?doid=2145204.2145354.

Kurasaki, K.S., 2000. Intercoder Reliability for Validating Conclusions Drawn from Open-Ended Interview Data. *Field Methods*, 12(3), pp.179–194.

Kuzon, W., Urbanchek, M.G. & Mccabe, S.J., 1997. Seven deadly sins of statistical analysis. *Journal of Oral and Maxillofacial Surgery*, 55(8), pp.897–898. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0278239197903773.

Lang, T. a & Altman, D.G., 2013. Basic Statistical Reporting for Articles Published in Biomedical Journals : The " Statistical Analyses and Methods in the Published

Literature " or The SAMPL Guidelines ". *Science editors' handbook*, pp.29–32. Available at: http://www.equator-network.org/wp-content/uploads/2013/07/SAMPL-Guidelines-6-27-13.pdf.

Langlois, R.N., 2002. Modularity in technology and organization. *Journal of Economic Behavior and Organization*, 49(1), pp.19–37.

Lee, J. & Lai, K.Y., 1991. What's in Design Rationale? *Humanï¿½Computer Interaction*, 6(3–4), pp.251–280.

Leek, J.T. & Peng, R.D., 2015. P values are just the tip of the iceberg. *Nature*, 520(7549), p.612.

Lukacs, P.M., Burnham, K.P. & Anderson, D.R., 2010. Model selection bias and Freedman's paradox. *Annals of the Institute of Statistical Mathematics*, 62(1), pp.117–125.

MacDonald, J., 2003. Assessing online collaborative learning: Process and product. *Computers and Education*, 40(4), pp.377–391.

MacLean, A. et al., 1991. Questions, Options, and Criteria: Elements of Design Space Analysis. *Humanï¿½Computer Interaction*, 6(3–4), pp.201–250.

Malone, T.W. et al., 1999. Tools for Inventing Organizations : Toward a Handbook of Organizational Processes Tools for Inventing Organizations : Toward a Handbook of Organizational Processes. , 3(May 2015), pp.425–443.

Mann, M., 2016. Must try harder. *New Scientist*. Available at: http://www.sciencedirect.com/science/article/pii/S0262407916303682 [Accessed August 30, 2016].

Martin Bland, J. & Altman, D., 1986. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet*, 327(8476), pp.307–310.

Mascha, E.J., 2010. Equivalence and noninferiority testing in anesthesiology research. *Anesthesiology*, 113(4), pp.779–781.

Miles, M., Huberman, M. & Saldana, J., 2014. *Qualitative Data Analysis*,

Morton, K. et al., 2014. Support the Data Enthusiast: Challenges for Next-Generation Data-Analysis Systems. *Proceedings of the VLDB Endowment, Volume 7, pp. 453–456, 2014*, 7, pp.453–456. Available at: http://homes.cs.washington.edu/~kmorton/p446-morton.pdf.

Nimon, K.F., 2012. Statistical assumptions of substantive analyses across the general linear model: A mini-review. *Frontiers in Psychology*, 3(AUG), pp.1–5.

Van Noorden, R., 2014. Online collaboration: Scientists and the social network. *Nature*, 512(7513), pp.126–129. Available at: http://www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711.

Norman, D.A., 1983. Some Observations on Mental Models. In *Mental Models*. pp. 7–14. Available at: http://www.amazon.com/Mental-Models-Cognitive-Science-Series/dp/0898592429.

Nosek, B.A., Spies, J.R. & Motyl, M., 2012. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, 7(6), pp.615–631. Available at:

http://pps.sagepub.com/lookup/doi/10.1177/1745691612459058.

Nuzzo, R., 2014. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487), pp.150–152.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251), p.aac4716-aac4716. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26315443.

Osborne, J. & Waters, E., 2002. Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research and Evaluation*, 8(2), p.1.

Ott, E.M., 1989. Effects of the Male-Female Ratio at Work: Policewomen and Male Nurses. *Psychology of Women Quarterly*, 13(1), pp.41–57.

Paglieri, F., 2004. Data-oriented belief revision: Towards a unified theory of epistemic processing. *Proceedings of STAIRS*. Available at: http://books.google.com/books?hl=en&lr=&id=Z569jqwQuK8C&oi=fnd&pg=PA179&dq=Data-oriented+Belief+Revision+:+Towards+a+Unified+Theory+of+Epistemic+Processing&ots=SqAEHHjdec&sig=Out0eaWHx3vygt_eBCzHm6QJYcA.

Partington, D., 2013. *Essential Skills for Management Research*,

Peffers, K. et al., 2008. A Design Science Research Methodology for Information Systems Research. *J. Manage. Inf. Syst.*, 24(January), pp.45–77.

Ransbotham, S., Kiron, D. & Prentice, P.K., 2015. The Talent Dividend. *MIT Sloan Management Review*, 56(4), pp.1–12. Available at: http://sloanreview.mit.edu/projects/analytics-talent-dividend/.

Redmiles, D., 2000. Software Requirements for Supporting Collaboration through Categories.

Reinecke, K. & Bernstein, A., 2013. Knowing What a User Likes: A Sesign Science Approach to Interfaces that Automatically Adapt to Culture. , 37(2), pp.427–453.

Rouder, J.N. et al., 2016. Is There A Free Lunch In Inference? *topiCS*, 8(1), pp.1–5.

Russell, D.M. et al., 1993. The cost structure of sensemaking. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '93*. pp. 269–276. Available at: http://portal.acm.org/citation.cfm?doid=169059.169209.

Russo, D. & Zou, J., 2016. Controlling Bias in Adaptive Data Analysis Using Information Theory. In *Proceedings of the 19th International Con- ference on Artificial Intelligence and Statistics (AISTATS) 2016*. pp. 1232–1240. Available at: http://arxiv.org/abs/1511.05219.

Saldana, J., 2011. *Fundamentals of Qualitative Research: Understanding Qualitative Research*,

Salehi, N. et al., 2016. Huddler: Convening Stable and Familiar Crowd Teams Despite Unpredictable Availability. *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

dos Santos, F. & Bazzan, A.L.C., 2009. An ant based algorithm for task allocation in large-scale and dynamic multiagent scenarios. *Proceedings of the 11th Annual conference on Genetic and evolutionary computation - GECCO '09*, p.73. Available at:

http://portal.acm.org/citation.cfm?doid=1569901.1569912.

Van Schaik, P. & Weston, M., 2016. Magnitude-based inference and its application in user research. *International Journal of Human Computer Studies*, 88(August), pp.38–50.

Schlauderer, S. & Overhage, S., 2013. Exploring the Customer Perspective of Agile Development: Acceptance Factors and on-Site Customer Perceptions in Scrum Projects. *Thirty Fourth International Conference on Information Systems*, pp.1–20.

Schubanz, M., 2014. Design rationale capture in software architecture: What has to be captured? In *WCOP 2014 - Proceedings of the 19th International Doctoral Symposium on Components and Architecture (Part of CompArch 2014)*. pp. 31–36. Available at: http://dx.doi.org/10.1145/2601328.2601329.

Sculley, D. & Pasanek, B.M., 2008. Meaning and mining: The impact of implicit assumptions in data mining for the humanities. *Literary and Linguistic Computing*, 23(4), pp.409–424.

Seel, N.M., 2001. Epistemology, situated cognition, and mental models: "Like a bridge over troubled water." *Instructional Science*, 29(4–5), pp.403–427.

Seitz, F., Heisenberg, W. & Pauli, W., 2000. Decline of the generalist The vigour of every discipline depends on people of broad vision . *Nature*, 403(February), pp.10021–10021.

Sere, F.C. et al., 2011. Online collaboration: Collaborative behavior patterns and factors affecting globally distributed team performance. *Computers in Human Behavior*, 27(1), pp.490–503.

Sheskin, D.J., 2004. Handbook of parametric and nonparametric statistical procedures. *Technometrics*, 46, p.1193. Available at: http://books.google.com/books?id=bmwhcJqq01cC&pgis=1.

Silberzahn, R. & Uhlmann, E.L., 2015. Many Hands Make Tight Work. *Nature*, 526(7572), pp.189–191. Available at: http://www.nature.com/news/crowdsourced-research-many-hands-make-tight-work-1.18508.

Simmons, J.P., Nelson, L.D. & Simonsohn, U., 2011. False-Positive Psychology. *Psychological Science*, 22(11), pp.1359–1366. Available at: http://journals.sagepub.com/doi/10.1177/0956797611417632.

Smith, A.J., 1990. The Task of the Refereee. , pp.1–7.

Stefik, M., 1981. Planning with constraints (MOLGEN: Part 1). *Artificial Intelligence*, 16(2), pp.111–139.

Stein, R.T. & Heller, T., 1979. An empirical analysis of the correlations between leadership status and participation rates reported in the literature. *Journal of Personality and Social Psychology*, 37(11), pp.1993–2002.

Strasak, A.M. et al., 2007. Statistical errors in medical research - A review of common pitfalls. *Swiss Medical Weekly*, 137(3–4), pp.44–49.

Strauss, A. & Corbin, J., 1990. Basics of qualitative research: grounded theory procedure and techniques. *Qualitative Sociology*, 13(1), pp.3–21.

Thomas, D.R., 2006. A General Inductive Approach for Analyzing Qualitative Evaluation Data. *American Journal of Evaluation*, 27(2), pp.237–246. Available at:

http://journals.sagepub.com/doi/10.1177/1098214005283748.

Tseng, H. et al., 2009. Key Factors in Online Collaboration and Their Relationship to Teamwork Satisfaction. *The Quarterly Review of Distance Education*, 10(626), pp.195–206.

Tukey, J.W. & Wilk, M.B., 1966. Data analysis and statistics: an expository overview. *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference*, (695), pp.695–710. Available at: http://dl.acm.org/citation.cfm?id=1464366.

Vargha, A. & Delaney, H., 1998. The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*, 23(2), pp.170–192.

Viegas, F.B. et al., 2007. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), pp.1121–1128.

Weiss, G. & Wodak, R., 2003. *Critical Discourse Analysis*,

Westfall, J. & Yarkoni, T., 2016. Statistically controlling for confounding constructs is harder than you think. *PloS one*, 11(3), pp.1–22.

Willett, W. et al., 2011. CommentSpace: Structured Support for Collaborative Visual Analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp.3131–3140. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.188.1845&rep=rep1&type=pdf.

de Winter, J.C.F. & Dodou, D., 2010. Five-Point Likert Items : t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11), pp.1–16. Available at: http://pareonline.net/pdf/v15n11.pdf.

Woolston, C., 2015. Psychology journal bans P values. *Nature*, 519(7541), pp.9–9. Available at: http://www.nature.com/doifinder/10.1038/519009f [Accessed August 12, 2017].

Yadav, M.S. & Pavlou, P.A., 2014. Marketing in Computer-Mediated Environments: Research Synthesis and New Directions. *Journal of Marketing*, 78(1), pp.20–40. Available at: http://journals.ama.org/doi/abs/10.1509/jm.12.0020.

Yukl, G., 2001. Leadership in organizations. *Personnel Psychology*, 7th(4), p.542. Available at: http://files.liderancaecoaching.webnode.com/200000015-31f5732fb3/media-F7B-97-randd-leaders-business-yukl.pdf.

Zimmerman, D.W., 2004. Inflation of Type I error rates by unequal variances associated with parametric, nonparametric, and rank-transformation tests. *Psicologica*, 25(1), pp.103–133.

Zimmerman, D.W., 1998. Invalidation of Parametric and Nonparametric Statistical Tests by Concurrent Violation of Two Assumptions. *The Journal of Experimental Education*, 67(1), pp.55–68.

8. Appendix

A1) Description of edge.org dataset

Explained variability for different cluster sizes:

Our dataset build started with collecting information from the Edge.org on all of the conversations and annual questions. We built a program that downloaded the information from the website, including the year, title, link to, and type of the

conversation, as well as the text itself and who said it. Two independent coders then coded gender of the contributors based on their profile picture on Edge.org, or, if that was not available, pictures and pronouns on other reputable websites. We then manually collected information on the job title, workplace, and PhD by finding CVs, university webpages, news articles, personal websites, and Linkedin profiles. We wrote a program to collect the US News and World Report International Rankings and the Shanghai Rankings and manually gathered the rankings from the National US News and World Report Rankings. We then ran the text through the LIWC program. Finally, we calculated the rest of the variables (such as male contributors, previous contributions, etc.) based on the data we had already collected.

The descriptions below include the name in the full version of the dataset and the shortened name used in the dataset for older software.

- Conversation level:
    - **Year**: the year when it took place
    - **Title:** the title of the conversation. For example: "What Scientific Idea Is Ready For Retirement?"
    - **Link:** a link to the conversation
    - **Type:** 1 for annual question, 2 for conversation
        - Edge does an **annual question** every year; some examples are "what scientific idea is ready for retirement?" and "What will change everything?" People then write in with their answers. So all of the text is written and asynchronous
        - What Edge refers to as a **conversation** can actually be multiple things. Some of these are written essays by a single person, some are transcripts of a speech, and some are transcripts of a conversation (either between two or more guests or an interview).
    - **ThreadID (ThrdID):** a unique identifier for each conversation/annual question (between two or more people)
    - **MaleContributions (Mcontr):** the **number of times** a man speaks in a specific conversation, it **does not** always equal the number of unique men in a conversation (see below)
    - **FemaleContributions (Fcontr):** the **number of times** a woman speaks in a specific conversation, it **does not** always equal the number of unique women in a conversation (see below)
    - **FemaleParticipation (Fpart):** simply femalecontributors/(number of total contributions); the percentage of comments that are made by a woman
    - NumberAuthors (NumAut):
        - For the annual questions, this equals 0; because the website is the author of the question, everyone is considered commentators
        - Otherwise, this is the total number of times people contribute to the main body of the text, rather than people who just comment. For example, in http://edge.org/conversation/how-democracy-works-or-why-perfect-elections-should-all-end-in-

ties, there are multiple people commenting on the post, but W. Daniel Hillis is the only author and only speaks once (as it is an essay). So NumberAuthors is "1." If two people each spoke five times in a dialogue, NumberAuthors would be "10."

- **DebateSize (DebSiz):** number of text pieces in a conversation; it is the sum of female and male contributions
- **Live:** whether the text piece was transcribed or written; it is 0 if it is written (either an essay or a comment on a piece) and 1 if it was part of a live conversation or speech that was later transcribed**.** Here are the types of text and how they would be classified:
  - **A single author essay** (live = 0 because it is written): http://edge.org/conversation/the-evolved-self-management-system
  - **A single author speech** (live = 1 because it was spoken and later transcribed): http://edge.org/conversation/cities-as-gardens
  - **A live conversation**, either between multiple people or in an interview format (live = 1 because it was spoken and later transcribed): http://edge.org/conversation/japan-inc-meets-the-digerati
  - **Online Comments** on any of the three types above (live = 0 because it was written)
  - **The annual question (Type = 1):** live = 0 because these were all written and submitted.
- UniqueContributors (UContr): UniqueMaleContributors + UniqueFemaleContributors
- **UniqueMaleContributors (UMContr)**: the number of unique male contributors
- **UniqueFemaleContributors (UFContr):** the number of unique female contributors
- UniqueFemaleParticipation (UFPar): the percentage of unique female participants; UniqueFemaleContributors divided by UniqueContributors

- Participant Level
  - **Id:** the unique identifier of the contributor
  - **Id_num**: the unique identifier of the contributor as text (this is typically the format of first name_last name)
  - **Role**: Either author (=1) or commentator (=2)
  - **Name**: name of the commentator
  - **TwoAuthors (TwoAutrs):** some of the edge comments are written by two people. In this case, we duplicated the row and kept the text level and conversation level information the same and had one author per row. This variable is 1 if this text was written by two people and 0 otherwise.
  - **Female**: the commentator is male = 0, the commentator is female = 1
  - **Male**: the commentator is female = 0; the commentator is male = 1
  - **Academic (Acad):** 1 = the person is in academia, 0 = they are not

- **Limited_Information (LimInfo):** equals 1 if we could only find limited information about the person (e.g. they commented in 2013 but we only have their job title from 2012), 0 otherwise
- **Job_Title (JobT)**: The job title of the commentator
- **Job_Title_S (JobTS):** This is a simplified list of job titles (e.g. we have "Eugene Higgs Professor" in Job.Title but "Chaired Professor" in Job.Title.Collapsed)
    - Chaired Professor
    - Professor
    - Associate Professor
    - Assistant Professor
    - Non-Tenure-Track Faculty
    - Postdoctoral Researcher
    - Graduate Student
    - Academic Leadership (Dean, Vice President, etc.)
    - Researcher
    - Artist/Author/Editor/Writer
    - Director
    - Founder
    - Other
    - Top Management and Founder
    - Top Management
    - Entrepreneur
    - Not Available
- **Job_Title_S_num (JobTSn):** Job_Title_S as numbers instead of text
- **Department (Dept):** what academic department someone is in
- **Department_S (DeptS):** a simplified version of all the departments (e.g. while John Smith's Department is "Experimental Physics," his Department_S is "Physics")
    - Physics (Phy)
    - Anthropology (Ant)
    - Earth Sciences (ES)
    - Biology (Bio)
    - Psychology (Psych)
    - Journalism, media studies and communication (JMS)
    - Medicine (Med)
    - Philosophy (Phil)
    - Space Sciences (SS)
    - Linguistics (Lin)
    - Computer Sciences (CS)
    - Engineering (Eng)
    - Arts (Arts)
    - Business/Management (Bus)
    - Environmental Studies and Forestry (ESF)
    - Sociology (Soc)
    - Mathematics (Math)
    - Asian Studies (AS)
    - Education (Educ)

    - Political Science (PS)
    - Economics (Econ)
    - Systems Science (Sys)
    - History (Hist)
    - Music (Musc)
    - Chemistry (Chem)
    - Archeology (Arch)
    - Architecture and Design (ArchD)
    - Law (Law)
    - Zoology (Zoo)
    - Literature (Lit)
    - Divinity (Div)
- **Department_S_num (DeptSn):** Department_S as numbers instead of text
- **Discipline (Disc)**: this groups academic departments into disciplines
    - Natural Sciences (NS)
    - Social Sciences (SocS)
    - Professions (Prof)
    - Humanities (Hum)
    - Formal Sciences (FS)
- **Workplace (Workpl):** where someone works; some people are self-employed
- **HavePhD (PhD):** equals 1 if they have a phd, 0 otherwise. It is 1 even if someone earns a phd after they comment (e.g. John Doe comments in 2000 and earns his PhD in 2012; his comment in 2000 will still have HavePhD = 1)
- **PhD_Field (PhDF):** what field people got their PhD in
- **PhD_Year (PhDY):** what year they got their PhD
- **PreviousContributions (PrContr)**: how many times **before this year** they have made contributions. So if John Doe only talked three times in one conversation in 2012 and one time each in two conversations in 2014 (and never made any other comments), this will be 0 for his comment in 2012 and 3 for both his comments in 2014.
- **ContributionsThisYear (ContrTY):** how many times they contributed this year; even if they only participated in one conversation, if they spoke 40 times in that conversation, this variable will be 40.
- **ThreadsThisYear (ThrTY):** how many threads they participated in this year; thus if John spoke in two threads in 2014, one twenty times and one once, this would equal 2 in 2014, while **ContributionsThisYear** would equal 21 for 2014.
- **PreviousThreads (PreThrd):** how many threads they participated in **before this year**. So, if John contributed for the first time twice in one thread in 2000, once each in two different threads in 2004, and once in 2014, this would be 0 for 2000, 1 for 2004, and 3 for 2014 (and for **PreviousContributions** it would be 0 for 2000, 2 for 2004, and 4 for 2014).

- **AuthorandCommentator (AutAndCom):**: if, for the same piece, someone is both an author and a commentator, this is 1 for that person for that piece; otherwise it is 0
- **PhD_Institution (PhDI)**: what school they got their PhD
- **Years_from_PhD (YfPhD):** how many years at the time of the comment since they earned their PhD; this is just Year - PhD.Year. This can be negative because people may have earned their phd years after they make a comment
- **PhD_Institution_SR (PhDISr):** The Shanghai Rankings of their PhD Institution; this is only for people who received their PhDs from institutions that are ranked by Shanghai. Shanghai ranks only between 500 and 510 universities worldwide each year and also bins their rankings after a certain point, in different ways for different years (e.g. a university may be ranked as 301-352).
- PhD_Institution_SR_Bin (PhDISrB):
    - 1 = university was ranked between 1 and 50
    - 2 = university was ranked between 51 and 100
    - 3 = university was ranked between 101 and 150
    - 4 = university was ranked between 151 and 200
    - 5 = university was ranked between 201 and 300
    - 6 = university was ranked between 301 and 400
    - 7= university was ranked between 401 and 510
- **Workplace_SR (WorkSr):** The Shanghai Rankings of their workplace; this is only for academics and academic institutions that are ranked by Shanghai (see **PhD_Institution_SR** for more information)
- Workplace_SR_Bin (WorkSrB):
    - 1 = university was ranked between 1 and 50
    - 2 = university was ranked between 51 and 100
    - 3 = university was ranked between 101 and 150
    - 4 = university was ranked between 151 and 200
    - 5 = university was ranked between 201 and 300
    - 6 = university was ranked between 301 and 400
    - 7= university was ranked between 401 and 510
- **SR_Ranking_Dif (SrRDif):** The difference between the binned Shanghai Ranking University of their workplace and the binned Shanghai Ranking of their PhD; a positive ranking means that they work at a place that has a higher ranking than where they got their PhD
- **PhD_Institution_US_IR (PhDIR):** The US News and World Report created an international ranking system in 2014 to rank the top 500 universities. Thus, even if a comment was made in 1999, if they have a PhD from Carnegie Mellon, this ranking will be Carnegie Mellon's ranking in the 2014 report
- PhD_Institution_US_IR_Bin (PhDIRB):
    - 1 = university was ranked between 1 and 50
    - 2 = university was ranked between 51 and 100
    - 3 = university was ranked between 101 and 150
    - 4 = university was ranked between 151 and 200

    - 5 = university was ranked between 201 and 250
    - 6 = university was ranked between 251 and 300
    - 7= university was ranked between 301 and 350
    - 8 = university was ranked between 351 and 400
    - 9 = university was ranked between 401 and 450
    - 10 = university was ranked between 451 and 500
- Workplace_US_IR (WorkIR): See PhD_Institution_US_IR
- Workplace_US_IR_Bin (WorkIRB):
    - 1 = university was ranked between 1 and 50
    - 2 = university was ranked between 51 and 100
    - 3 = university was ranked between 101 and 150
    - 4 = university was ranked between 151 and 200
    - 5 = university was ranked between 201 and 250
    - 6 = university was ranked between 251 and 300
    - 7= university was ranked between 301 and 350
    - 8 = university was ranked between 351 and 400
    - 9 = university was ranked between 401 and 450
    - 10 = university was ranked between 451 and 500
- **USA_I_Ranking_Dif (IRDif):** the difference between the rank of someone's workplace and the rank of their PhD Institution (as ranked by US News and World Report International Rankings). If this is positive, it means they're working at an institution ranked higher than their PhD Institution.
- **PhD_Institution_US (PhDIUS):** The ranking of their PhD Institution by USA News and World Report; this is **only** for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These only go from 2003-2014.
- PhD_Institution_US_Bin (PhDIUSB):
    - 1 = university was ranked between 1-5
    - 2 = university was ranked between 6-10
    - 3 = university was ranked between 11-25
    - 4 = university was ranked between 26-50
    - 5 = university was ranked between 51-100
    - 6 = university was ranked between 101-150
    - 7 = university was ranked between 151-200
- **Workplace_US (WorkUS):** The ranking of their workplace by USA News and World Report; this is **only** for US institutions and only for a limited number of them. Different numbers of school were ranked in different years; for example, 129 schools were ranked in 2005, while only 51 were ranked in 2003. These only go from 2003-2014.
- Workplace_US_Bin (WorkUSB):
    - 1 = university was ranked between 1-5
    - 2 = university was ranked between 6-10
    - 3 = university was ranked between 11-25
    - 4 = university was ranked between 26-50
    - 5 = university was ranked between 51-100
    - 6 = university was ranked between 101-150

        - 7 = university was ranked between 151-200
    - **USA_Ranking_Dif (USRDif):** the difference between the rank of someone's workplace and the rank of their PhD Institution (as ranked by US News and World Report Rankings). If this is positive, it means they're working at an institution ranked higher than their PhD Institution.
    - **Total_Citations (TotCit):** the total number of citations they have received, including that year and all previous years (it's citations.year + previouscitations)
    - **H_Index (Hind):** this is their h-index in **2014;** a scholar has an index of *h* if they have published *h* papers each of which has been cited in other papers at least *h* times
    - **i10_index (iTenIn):** how many papers in **2014** they had authored that has more than 10 citations; this is only for Google Scholar pages. As the GS pages only have an i10 index from 2014, even if the comment was from 1999, the i10 index is from 2014
    - **Citations_Year (CitY):** how many citations they received this year; this is only for Google Scholar pages, so not all academics have this
    - **Citations_Cumulative (CitCum):** how many citations they have received in this year and previous years; this is only for Google Scholar pages, so not all academics have this
    - AcademicHierarchyStrict (AcaHier):
        - 1 = Graduate Student
        - 2 = Postdoctoral
        - 3 = Assistant Professor
        - 4 = Associate Professor
        - 5 = Professor
        - 6 = Chaired Professor
    - **PreviousCitations (PreCit):** the number of citations they have received in all of the previous years
    - **ContributionsbyAuthor (ContrAut):** the number of contributions by this author in this conversation
    - Dummy variables for Discipline
    - Dummy variables for department_S
- Text-Level
    - **Order:** The order of the text pieces. This is **meaningless** for Annual Questions.
    - **Text:** the text of the conversation
    - **Number_Characters:** number of characters in the text piece
    - LIWC variables (see www.liwc.net/descriptiontable1.php)

Part III, Epilogue

Epilogue

Part II of this thesis included four articles exploring potential ways to contribute to data science through the crowdsourcing lens. We first presented a use case of laymen crowd workers contributing to the review process of statistical reporting in research papers. This use case demonstrated that with a properly designed process even crowds with no expertise in statistics can contribute to ease some of the burning scientific issues. This contribution can be seen as another opportunity for citizen science where everyone can contribute to the scientific progress or, in our case, scientific improvement.

We then conducted an exploratory study to better understand what it takes to outsource some tasks in data science project to freelancers available on OLMs. This study led to a reconfirmed understanding that various bottle-neck activities of data preprocessing could be potentially crowdsourced. Guided by this realization, we proposed a use case where non-experts collaborate online with data scientists to perform various data analysis tasks. For this, we designed a prototype of the platform that would facilitate such collaboration and provided evidence that the results of such setting might be of high quality and competitive cost.

Our final contribution was a study where we explore the factors underlying the variability in data analysis and driving data analysts reach different results while conducting the same data analysis.

All of these contributions are subject to limitations. In the next section, we will outline general limitations and directions for future work. Note that more specific limitations have been elaborated in their corresponding articles in Part II of this dissertation.

Limitations and Future Work

This section outlines some of the most important limitations of the methods and results presented in this dissertation. More details for each individual project can be found in the respective sections in Part II of this thesis.

Reliability

The survey studies we conducted both in the second and the third studies had a limited sample size. Our interview study with data scientists was geographically limited to Switzerland and Germany and included 20 data scientists. The survey of the non-experts regarding the perceived complexity of the derived sub-tasks was limited to dozen respondents. Such small sample sizes naturally jeopardize the reliability of our outcomes since too small n reduces the statistical power of the study and increases the margin of error. Also, the first paper, presenting cross-disciplinary comparison on the statistical assumption reporting practices, lacking enough observations in order to statistically evaluate the difference between fields. However, all these three articles did not reach the required sample size due to the scope and complexity of the studied phenomena and the high cost which would be required to achieve the necessary level of analysis power. In the tradeoff between the ability to reach statistical significance and the challenge of researching a subject that is difficult to fully address from the strict statistical perspective due to high costs, we

decided in favor of the latter. Also, in our last research question, even though the theoretical saturation was reached while generating the system of codes and the corresponding categories, one can consider further quantitative analysis to verify our results. Again, due to the complexity of the studied phenomena we leave quantitative evaluation out of the scope of this study and propose it as a future research.

Experiment design

In our first article we performed experiment with CrowdStat to evaluate how well can crowds evaluate statistical assumption reporting. Crowdstat is designed in such way that by answering number of questions crowds indirectly estimate whether the assumption was reported. However, this approach has a limitation since some assumptions have natural prevalence in English and can be easily confused by crowds. For example, words like "normal" or "independent" might relate not only to statistical assumptions but also occasionally be used in another context. In its current form our approach does not offer a way to distinguish between the two cases. Additionally, our assumption is that crowds have no statistical knowledge. However, we did not obtain this information through survey.

The experiment we conducted in the third study, where we explored whether the data preprocessing could be outsourced to non-experts, had a few limitations related to the experiment design. For instance, we did not validate whether the top-down approach is necessarily the optimal process. We limited our study to such setting in order to simplify the experiment and having in mind that hierarchical task distribution is the most intuitive project structure and therefore would be easy to learn by our experiment participants.

In our last article, which is discussing the variability in data analysis, we asked analysts to annotate executed code after every 30-50 executed logs. This experimental design might introduce a natural bias as such labor intense task would be more thoroughly performed by analysts aware of the importance of these annotations and by those who have patience for such task. Additionally, by introducing code annotation after certain number of commands is executed we might interfere by changing the analysis style of analysts. Some analysts work in iterative manner where after every minor change they re-execute the whole code to ensure code compatibility. Therefore, re-executing all code again and again will create a frustrating number of requests to annotate the same code and might lead to sparse annotation.

Self-reporting

Our second research question is including survey where freelancers self-report their skills. Even though the respondents were instructed that the survey is only for research purposes, the results are likely to be positively biased to certain extent. This assumption is supported by the fact that workers are employed through Online Labor Markets based on the skills they possess, and therefore it would be only natural that they will overestimate their talents and skills. On the other hand, we took this inherent bias into consideration during our study and resulted with very conservative conclusion assuming that most of them have basic coding skills that will allow them to perform data preprocessing given that they will receive accurate instructions.

In the follow-up study we asked the participants to self-report on the complexity of the decomposed tasks to estimate to what extend the task was simplified. Here too we can not be entirely sure that the reported results were unbiased.

Conclusion

Given the technological trends of recent years, the demand for experts in the field of data analysis is likely to keep growing. Although universities have stepped up their efforts to close this gap, it appears that the shortage of skilled data scientists will remain for at least the next five-ten years. At the same time, there are more and more educated employees who prefer to avoid a restrictive setting of regular work and instead seek flexible employment frameworks as freelancers. In my dissertation, I presented scenarios of how to tap into the existing talent available online to satisfy the needs of industry in the data science domain. As I progressed in my research, it became clear that the present body of research proposes solid theoretical foundations of crowdsourcing and online collaboration, but to a certain extent still lacks an applied research that explores how to translate existing knowledge into practical prototypes of tools to support data analysis with diverse crowds. This was addressed through three studies where we proposed and evaluated scenarios of crowds' assistance to the data science process. The exploration was divided into three parts according to the degree of crowds' skills.

The first study demonstrated how laymen crowds without any expertise in the data analysis could solve a reasonable problem of evaluating the reporting of statistical assumptions in research papers. Such reliable, cheap and quick method for evaluating statistical assumption reporting is important as it I) allows to reduce the workload of scientific reviewers, II) has the potential to engage citizen into scientific process and, III) encourages further research on crowdsourcing other aspects of scientific quality assurance to the broad public. Moreover, besides the use-case, the developed method allowed for comprehensive meta-analysis of hundreds of papers at low cost and with quality equivalent to experts. We were able to analyze corpus of papers from different journals and disciplines and obtained a high-level understanding on statistical assumption reporting across different fields.

Our second and third study demonstrated how non-experts could be instrumental in data science process. Since we wanted to get a better insight on what skills freelancers possess, we also conducted a survey study where we learned about the skills of freelancers active on Online Labor Markets. After studying the tasks that expert data scientists would like to crowdsource and the skills of freelancers, we concluded that it would be possible to outsource tasks that require certain coding skills. Specifically, if provided with appropriate coordination system, data scientists would like to outsource data preprocessing activities which take lion's share of this time but do not require any advanced skills besides clear understanding on the desired shape of the data. This study resulted with case study demonstrating how experts (i.e. data scientists) and non-experts could collaborate together in a cost-efficient manner. As to the privacy issue we left it out of scope of this study as there are research fields set on a mission to solve this problem and enable unhindered data analysis of private data (cf. research on differential privacy).

Our fourth study revolved around crowdsourcing data analysis with experts. In this case the question was formulated differently. Given experts (even if they are rare), is there a scenario in which crowdsourcing could contribute to solving a challenge of

data science? As a result, we addressed a problem, which recently attracted attention of scientific community in the light of the crisis of confidence in science: the difficulty to replicate data-driven research. Specifically, even when the data and postulated hypothesis are identical, researchers often reach different results in their data analysis. By crowdsourcing the same data analysis to multiple researchers, we were able to propose factors responsible for the variability in data analysis. Moreover, the designed platform DataExplained offers means for more transparent data analysis through careful tracking of rational for decisions made throughout data analysis.

Combined, the four articles described here add up to a dissertation which is drawing on crowdsourcing as well as data science and presents alternatives to engage broader population in the process of data analysis. By doing so two major benefits are achieved: i) the growing shortage in data analysis experts might be to certain extent mitigated and ii) a broader audience might be engaged in data analysis and thus further foster the democratization of research by lowering the threshold for participation in data driven research.

University of Zurich[UZH]

Department of Informatics

University of Zurich
Department of Informatics
Binzmühlestr. 14
CH-8050 Zurich
Phone +41 44 635 43 11
Fax +41 44 635 68 09
www.ifi.uzh.ch

Bibliographic citations of the papers constituting this thesis

Paper 1: Assessing Statistical Assumption Reporting in CHI and Other Fields

This paper has been submitted to the ACM Transactions on Computer-Human Interaction journal

Paper 2: Data Analytics on Online Labor Markets: Opportunities and Challenges

This paper is published as *Feldman Michael, Frida Juldaschewa, and Abraham Bernstein. "Data Analytics on Online Labor Markets: Opportunities and Challenges." arXiv preprint arXiv:1707.01790, 2017*

Paper 3: Towards Collaborative Data Analysis with Diverse Crowds – a design science approach

This paper is published as *Feldman Michael, Cristian Anastasiu, and Abraham Bernstein. "Towards Collaborative Data Analysis with Diverse Crowds–A Design Science Approach." In International Conference on Design Science Research in Information Systems and Technology, pp. 218-235, 2018.*

Paper 4: Analysis of Behavioral Factors Underlying the Data Analysis Process

This paper will be submitted to a journal in the field of data science or psychology